Multi-Sensor Noise Suppression and Bandwidth Extension for Enhancement of Speech

A Dissertation
Presented to
The Academic Faculty

by

Rongqiang Hu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
School of Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
May 2006

Multi-Sensor Noise Suppression and Bandwidth Extension for Enhancement of Speech

Approved by:

Dr. David V. Anderson, Advisor
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Gordon Stuber
Professor, School of ECE
Georgia Institute of Technology

Dr. Mark Clements
Professor, School of ECE
Georgia Institute of Technology

Dr. Xu-Yan Chen
Associate Professor, Department of Math
Georgia Institute of Technology

Dr. Chin-Hui Lee
Professor, School of ECE
Georgia Institute of Technology

Date Approved: December 02, 2006

To my dear family, friends and my teachers

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. David Anderson, for the guidance, support, endurance, and encouragement during my graduate studies at Georgia Tech. He has always been supportive of me not only when I made progress but also when I made mistakes. I am also deeply impressed with his dedication to research and students. What I have learned from him are by no means limited to signal processing techniques, but rather applicable to all the life. His positive influence will no doubt propagate beyond one graduate degree and serve me well for a long time.

I also would like to thank Dr. Bhiksha Raj, who was my host at Mitsubishi Electric Research Lab. My knowledge was enhanced and my horizon was expanded after each technical conversation with him. I have also been greatly benefited from the interactions with Dr. Mark Clements, and I am thankful to his guidance.

More than four years of graduate studies at Georgia Tech provided me with the wonderful opportunity to work closely with several colleagues and co-researchers. I cherish the fruitful interactions I had with the researchers that I worked with on the DARPA supported low bit-rate speech coding research project. I also thank all my colleagues in the Efficient Signal Processing (ESP) Lab and Co-operative Analog and Digital Signal Processing (CADSP) group for their collaboration in several research adventures.

I am also thankful to my friends outside Georgia Tech for their encouragement during the past four years. My family members mean a lot to me. Especially, I am grateful to my eldest brother for his moral support. I also thank the fellows at Mitsubishi Electric Research Lab.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Speech enhancement has been an active research problem for decades and continues to be an important problem. This is made even more true by the proliferation of portable devices having audio input capabilities. In the presence of noise, both the quality and intelligibility of speech signals have been significantly deteriorated. Thus, to increase the performance of a speech processing system in real-world applications, a multi-sensor noise suppression system is proposed. Important side information from state-of-the-art non-air conductive sensors is utilized to drive a biologically inspired noise suppression algorithm and a glottal correlation filter in removing background noise. Also proposed is a speech bandwidth extension system that is used to extend to the wideband speech from the degraded speech acquired in the real-world, where the speech bandwidth is corrupted by background noise or transmission media.

Human hearing is a complicated non-linear procedure. Mathematically optimal solutions are not enough to achieve the requirement of subjective quality. Also, the statistical assumptions of the speech signal do not always hold true. The motivation of a biologically inspired Contant-Q (CQ) algorithm is to model the peripheral of the human auditory system and derive a perceptually optimal solution for noise suppression. In the proposed method, a detector for speech saliency is derived to measure the degree of conspicuousness of "significant" speech signals with the presence of background noise. Three biological correlates are exploited, including perception saliency determining the speech cues by mutual information between the energy in critical bands and speech/noise identification, phoneme saliency indicating the attributes in the speech production mechanism, and audibility saliency illuminating psychoacoustic masking properties by the frequency-to-place transformation of the basilar membrane. The detector generates frequency-based soft-decisions that are used in determining the presence of speech cues and controlling the parameters of speech signal processing to preserve interested components.

Another opportunity in speech enhancement is the development of alternative non-air conductive sensors. Two state-of-the-art non-air conductive sensors will be used. These are a electromagnetic sensor that can detect the internal motions of the glottis (GEMS) and a gel-embedded accelerometer (P-mic) that picks up vibrations of the skin associated with speech. The signals from each exhibit significant attenuations on ambient noise, but provide incomplete information about the speech signal. From the acquired signals, important side information, such as high-resolution segmentation, pitch epoches, and approximated glottal excitation can be obtained to enhance noise suppression algorithms and other algorithms. Since the GEMS signal is statistically correlated to the speech and independent of the ambient noise, an enhanced glottal correlation (GCORR) filter is developed and effective in cleaning up the noisy speech. Detailed analysis shows that the outputs of the two algorithms have different spectral characteristics. A hybrid multi-sensor system (CQ-GCORR) is derived to combine the strengths of these two algorithms. The fusion criterion is based on the maximum signal-to-noise ratio (SNR) improvement of the fused signals weighted by an *a* priori knowledge.

In many cases of the transmission for speech signals, such as a telephone system and low-bit-rate vocoder, the limitation of reliable bandwidth is a bottleneck for improving the quality and intelligibility of the degraded speech. This can occurs for the speech acquired in high noise environments or telephony, where the speech bandwidth is limited to 4KHz. In such cases. it is often useful to restore high frequency components. Therefore, the extension of bandwidth is of great importance in speech enhancement. A speech bandwidth extension algorithm is proposed using an improved codebook mapping method. The application of this algorithm is for the extension of telephony speech.

The proposed research are the frameworks for improving the quality/intelligibility of the degraded speech:

- a single-channel noise suppression system based on perceptual speech detection
- a multi-sensor noise suppression system for acoustic harsh environments based on non-air conductive sensors

- a speech bandwidth extension system for telephone speech

Extensive experiments indicate the significant improvement in both speech intelligibility and quality from the proposed frameworks.

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

Speech is the most used means of humans' communication. For over 100 years, the objectives in this area have been intensively studied. From the advent of the telephone to wireless communication networks, many bandwidth-efficient and high-quality transmission systems have been designed and have experienced explosive growth. The work and investigations on speech technology have been largely driven by real-world applications, which now have broaden to include not only communication or telephony, but also speech enhancement systems, efficient speech coding systems, automatic speech recognition systems, speaker verification systems, and speech modification systems. The objectives of much of the research have been to design and implement real-world workable and economically affordable systems that can be used over the existing and newly installed communication channels.

Speech signals are generally acquired by acoustic devices, such as acoustic microphones. This kind of measurement is susceptible to background noise, which exists in most real-world situations, such as car noise on the road, subway noise in underground transportation, and fan noise from air-conditioning at home. Therefore, it is often essential to process speech acquired in environments having high ambient white or colored noise. This is true for consumers (as in cell phone usage), industry (e.g. communications in factory environments), and the military. There are two principal perceptual criteria for measuring the performance of a speech system: quality and intelligibility. The quality of the enhanced signal measures its clarity, distorted nature, and the level of residual noise in that signal. The quality is a subjective measure that is indicative of the extent to which the listener is comfortable with the enhanced signal. The second criterion measures the intelligibility of the enhanced signal. This is an objective measure that provides the percentage of words that can be correctly identified by listeners. The presence of noise causes deterioration in both the quality and

the intelligibility of speech. Therefore, speech enhancement is an important and necessary tool to make speech systems workable in the real world.

In many speech transmission systems, such as the digital public telephone system, low-bit-rate-speech coding environments, the bandwidth of speech is limited to 4KHz. This kind of speech is so called "telephone speech." Compared to natural speech, telephone speech has a significantly degraded performance. The bandwidth limitation of telephone speech reduces speech intelligibility by about 10 percent, and decreases the subjective quality score, which is measure in terms of the subjective mean opinion score (MOS) by more than one point [56]. Owing to the importance of the acoustic bandwidth for speech intelligibility and especially for subjective quality, it is worthwhile to extend the speech bandwidth. Particularly, in digital communication and hands-free telephony, there is a demand for enhancing the subjective speech quality. Especially when high quality telecommunications services grows, the various communication networks are rapidly merging together, *e.g.*, merging wireline packet networks and wireless networks. The trend causes an inter-operability problem due to the different standards of each network. One of the important cases is that merging of the networks with different speech bandwidth, such as the connection from the public switched telephone network (PSTN) to wideband speech codec in voice over IP (VoIP) network, where results in significant quality degradation. This problem is one of the focuses in the cross-tandeming methods.

1.2 Contribution of the Thesis

This thesis focuses on speech enhancement in various real-world applications. The following approaches are introduced in this thesis:

- A recursive noise estimator based on the perceptual detection of voice
- Single-channel noise suppression rules using biologically inspired techniques based on the properties of the phoneme distinction and psychoacoustic masking
- A glottal correlation filter using non-acoustic devices, a glottal electromagnetic movement sensor (GEMS) and a physiology microphone (p-mic).

- A robust voice activity detector based on the facial movement detection from an ultrasonic Doppler sensor
- A speech bandwidth extension algorithm using improved codebook mapping towards increased phonetic classification

1.3 Organization of the Thesis

The thesis is organized as follows: Chapter 2 presents the background materials on the research in noise estimation, noise suppression rule, speech processing using non-acoustic sensors, and the bandwidth extension of speech. In Chapter 3, a perceptual voice detection based noise estimator and biologically inspired noise suppression rules are presented. Chapter 4 focuses on the description of a glottal correlation filter using state-of-the-art non-acoustic sensors and the integration of these techniques. A speech bandwidth extension algorithm for speech enhancement is presented in Chapter 5.Finally, the concluding remarks are given in Chapter 6.

CHAPTER 2

BACKGROUND

In terms of individual communication, speech is probably the most important and efficient means, even in today's multi-media society. Experimental tests by Ochsman and Chapanis [74] show that, as would be expected, the performance time of cooperative tasks performed in groups was up to ten times faster when speech was allowed compared to when it was not. Thus, with many rooms being used solely for speech between individuals and groups, it is important that acoustic designs accommodate and enhance such use. In real-world applications, the human speech is easily degraded by the background noise.

2.1 Noise Suppression

Within every acoustic environment, there is always a certain level of ambient background noise present. The level of this is mostly dependant on the activities taking place within the space and its more immediate surrounds. The most obvious effect of background noise is that it masks the speech signal, thus reducing the signal to noise ratio (SNR) as the receiver must specifically concentrate on the speech. We are able to understand speech in a moderately noisy environment because speech is a highly redundant signal and thus even if part of the speech signal is masked by noise, other parts of the speech signal will convey sufficient information to make the speech intelligible, or at least sufficiently intelligible to allow for effective speech communication. There is less redundancy in the speech signal for a person with hearing loss since part of the speech is either not audible or is severely distorted because of the hearing loss. Background noise that masks even a small portion of the remaining, impoverished speech signal will degrade clarity significantly because there is less redundancy available to compensate for the masking effects of the noise [6].

The design of noise suppression algorithms consists of two parts: noise estimation and suppression rule. The noise parameters, such as power spectrum, are estimated based on

statistical models about speech and noise signals. The estimates are then input into the suppression rules derived from the mathematical measures that are somehow believed to be correlated with the quality and/or intelligibility of the speech signal [23, 20, 24, 25].

2.1.1 Noise Estimation

The estimation of the characteristics of noise [69, 70, 77] in acoustic speech signals is essential for noise suppression algorithms. Traditional noise estimation methods, which are based on voice activity detector (VAD), restrict the update of the estimate to periods of speech absence. Additionally, VAD is generally difficult to tune and their reliability severely deteriorates for weak speech components and low input SNR [65, 67, 84]. Alternative techniques, based on histograms in the power spectral domain [39, 64, 81], are computationally expensive,require much memory resources, and do not perform well in low SNR conditions. Furthermore, the signal segments used for building the histograms are typically of several hundred milliseconds, and thus the update rate of the noise estimate is essentially moderate.

A useful noise estimation approach, known as the minimum statistics (MS) [62], is to track the minima of a smoothed power estimate of the noisy signal, and multiply the result by a factor that compensates the bias. However, the variance of this noise estimate is about twice as large as the variance of a conventional noise estimator [62]. Moreover, this method may occasionally attenuate low energy phonemes, particularly if the minimum search window is too short [14]. These limitations can be overcome, at the price of significantly higher complexity, by adapting the smoothing parameter and the bias compensation factor in time and frequency [63]. A computationally more efficient minimum tracking scheme is presented in [17]. Its main drawbacks are the very slow update rate of the noise estimate in case of a sudden rise in the noise energy level, and its tendency to cancel the signal [67]. Other closely related techniques are the lower-energy envelope tracking [81] and the quantile based [86] estimation methods. Rather than picking the minima values of a smoothed periodogram, the noise is estimated based on a temporal quantile of a non-smoothed periodogram of the noisy signal. Unfortunately, these methods suffer from the high computational complexity associated with the sorting operation,and the extra memory required for keeping past

spectral power values.

Recently, a noise estimation approach is introduced, namely minima controlled recursive averaging (MCRA) [13, 14], that combines the robustness of the minimum tracking with the simplicity of the recursive averaging. The noise estimate is obtained by averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability in subbands. The speech presence probability is controlled by the minima values of a smoothed periodogram. In contrast to the MS and related methods, the minimum tracking is not crucial,since it only controls the recursive averaging as a secondary procedure. The recursive averaging is carried out without a hard distinction between speech absence and presence, thus continuously updating the noise estimate even during weak speech activity. Additionally, the smoothing of the noisy periodogram is carried out in both time and frequency, which takes into ac-count the strong correlation of speech presence in neighboring frequency bins of consecutive frames. The MCRA noise estimate is shown computationally efficient, and characterized by the ability to quickly follow abrupt changes in the noise spectrum. Further improvement for the MCRA estimator includes the following aspects: Minimum tracking during speech activity, speech presence probability estimation, and derivation of a bias compensation factor. The proposed procedure comprises two iterations of smoothing and minimum. tracking. The first iteration provides a rough voice activity detection in each frequency band. Then, the smoothing in the second iteration excludes relatively strong speech components,which makes the minimum tracking during speech activity robust. This facilitates larger smoothing windows, and thus a decreased variance of the minima values. The estimation of the speech presence probability is based on a Gaussian statistical model [22]. However, the a priori speech absence probability is controlled by the result of the minimum tracking. This prevents the estimated noise from increasing during weak speech activity, especially when the input SNR is low. The speech presence probability is biased toward higher values to avoid speech distortions in speech enhancement applications, accordingly, a factor to compensate its bias in the noise estimator. The value of the bias compensation factor is determined by the *a* priori speech absence probability estimator, and an explicit expression is derived.

These kinds of algorithms have the ability to update noise during speech activity. the detection of speech is characterized by soft-decision speech presence probability. The detection is based on the statistical assumption that the real and imaginary part of a Fourier transform coefficients can be considered to be independent and can be modeled as zero mean Gaussian variables [4]. However, it is pointed out that this assumption holds only when speech stationary with a relatively small span of correlation and for a large frame size. The statistical based speech detection is therefore ineffective in practical applications. The estimated noise variance may be too large, and occasionally attenuate low energy speech cues. The other drawback of these algorithms is that the estimation of speech presence probability is smoothed in frequency domain using only a small number of neighboring frequency bins. Actually, there exists certain correlation of all the critical sub-bands in the discrimination of different sounds [92]. In noise conditions, the problem is considered to be the discrimination of a speech cue and simultaneous acoustic disturbances. Therefore, the mutual information dependency in critical sub-bands can be used for detecting low speech cues.

2.1.2 Suppression Rule

Development of the first adaptive noise suppression rule is the spectral subtraction, which started from 1970's. Spectral Subtraction and other single microphone-based techniques estimate the frequency content of ambient noise and then attenuate it. While this approach provides some improvement, it has fundamental limitations.For example, if the noise level is equivalent to the voice level at a given frequency the noise cannot be attenuated without severely distorting the speech.

Both the speech quality and intelligibility are elaborate and expensive to measure, since they require listening sessions with live subjects. Thus, researchers often resort to less formal listening tests to assess the quality of an enhanced signal, and they use automatic speech recognition tests to assess the intelligibility of that signal. Quality and intelligibility are also hard to be quantified and expressed in a closed form which is amenable to mathematical optimization. Thus, the design of speech enhancement systems is often based

on mathematical measures that are somehow believed to be correlated with the quality and/or intelligibility of the speech signal. A popular example involves estimation of the clean signal by minimizing the mean square error (MSE) between the logarithms of the spectra of the original and estimated signals [21]. This criterion is believed to be more perceptually meaningful than the minimization of the MSE between the original and estimated signal waveforms [36, 57]. However, speech signals are perceived by the non-linear auditory system.

The difficulty in designing these kinds of noise suppression rules is the lack of explicit statistical models for the speech signal and noise process. In addition, the speech signal, and possibly also the noise process, are not strictly stationary processes. Common parametric models for speech signals, such as an autoregressive process for short-term modeling of the signal, and a hidden Markov process (HMP) for long-term modeling of the signal, have not provided adequate models for speech enhancement applications. A variant of the expectation-maximization (EM) algorithm, for maximum likelihood (ML) estimation of the autoregressive parameter from a noisy signal, was developed by Lim and Oppenheim [57] and tested in speech enhancement. Several estimation schemes, which are based on hidden Markov modeling of the clean speech signal and of the noise process, were developed over the years, see, e.g., Ephraim [21]. In each case, the HMP's for the speech signal and noise process were designed from training sequences from the two processes, respectively. While autoregressive and hidden Markov models have been proved extremely useful in coding and recognition of clean speech signals, respectively, they were not found to be sufficiently refined models for speech enhancement applications.

Biologically inspired noise suppression exploits the quantitative correlation of acoustical stimuli and human hearing sensations. Extensive results of psychoacoustic facts and models based on experimental data show that the important psychoacoustic elements used in speech enhancement are follows: frequency importance analysis, speech state dependency and psychoacoustic masking effects.

The critical-band concept, which describes the frequency separation of the perceptual significance of acoustic stimuli, has been widely used for speech enhancement and coding

[80]. In spectral substraction approaches [91], the frequency bins are divided into critical subbands and the speech signal envelopes are estimated separately in each band. But the mutual information dependency of each bands has not been exploited yet.

Speech state dependency, which can be referred to as the distinction of different phonemes in the human auditory system, is important for speech enhancement, since the influence of noise on each phoneme is different [80]. Drucker [18] has investigated five phoneme classes for distinct speech enhancement strategies. Kim [52] also introduced a modified version of spectral substraction using speech state input.

Psychoacoustic masking effects, the occlusion of one sound by another loud sound, is another important factor in hearing. This may occur if the sounds are simultaneous, or a loud sound can obliterate a sound closely following or preceding it. The basic idea of taking masking effects into account in speech enhancement is to attempt to make inaudible spectral components of annoying background residual by forcing them to fall below a masking threshold curve as derived from a measured speech spectrum. In previous speech enhancement algorithms, the efforts have only been made on spectral masking to control the enhancement gains [91]. However, temporal masking can not be ignored in noisy environment, since the audibility levels of speech signals in those regions are decreased significantly by the noise over-subtraction factors.

2.2 Secondary Sensors

It is known that the speech acquired by acoustic microphone is easily corrupted by the ambient noise. Today's technology has opened whole new realms in transducer technology. Ear microphones, Tooth microphones, Bone Conduction transducers, Electrical and Micro-Radar based transducers are but a few of the devices available to the designers of communication systems. Each brings its own benefits and drawbacks. The signals measured from these non-air conductive sensors exhibit significant attenuations on ambient noise, but provide incomplete information about the speech signal. The acquired side information can benefit the effort of speech enhancement, especially in low SNR cases.

Recently, a high-speed digital image recording of the vocal fold vibration at a rate of 4500

frames per second with the 256x256 picture elements has become possible. By using this technique, pattern of the vocal fold vibration under several different prosodic conditions, especially that of the utterance final position, were investigated. Imaging of the vocal fold vibration revealed that during the production of the vocal fry , the closure period from time to time become very lengthy. After this lengthy closure period,the vibration tends to start with weak oscillation and during the following cycles, the oscillation builds up. It is suggested that this process explains the occurrence of multiple vibratory patterns associated with the vocal fry.

Bone conduction is a process by which sound propagates through the skull. A bone conduction microphone, properly anchored on a persons head (typically at the mastoid process behind the year or the forehead) can transmit sounds generated by a person to another location. The advantages of such a device are severalfold, including that only speech made by the person, not the ambient noise, is picked up by the microphone. Such a device has many applications, including: space, manufacturing plants, skiing slopes, hiking, fire departments, etc. The device is also applicable for people who are bad at hearing or who may have had their outer ear damaged.

Newly developed electromagnetic (EM) near field sensors [71] (refer to the technology transfer web site) provide a capability for measuring EM wave reflections from speech organ interfaces in a non invasive, safe, fast, portable, and low cost fashion. These devices have similarities to some far-field radar systems except that their power levels are very low, their measurements are commonly in the near-field, and their rate of data acquisition is very high or very flexible. They are being used in investigations for many applications such as heart function and mechanical vibration sensing. In particular, they make possible the real-time measurements of the positions and motions of human vocal articulators during speech production. The measurements to date include the motions of the glottis (i.e., vocal folds), lips, tongue, jaw, and velum. Examples of vocal fold measurements (in real time) give the pitch period, enable noise removal, and realize pitch synchronous deconvolving of the corresponding excitation from the acoustic output to obtain quality-improved speech transfer functions. Similarly, EM sensor measurements of jaw motion with acoustic speech

provide constraints on the sound being articulated for speech recognition, and can be used for "talking head" and video image synchronization.

The physiological microphone (P-mic) is composed of a gel-filled chamber and a piezo-electric sensor behind the chamber [82]. P-mic measures the signal in response to applied forces that are generated by the movement of corresponding tissue where the sensor is placed. Because of the poor coupling properties between ambient noise and the fluid-filled pad, the output exhibits large attenuation of background noise. The P-mic signal at the throat contains clearer low-pass vocal tract formants than those of normal microphone.

The GEMS device and the P-mic provide two methods of extracting information about the glottal flow in high acoustic noise situations. This information can be used in a variety of ways for improving speech coding both in silence and in noise.

2.2.1 Voicing Detection

The first and most obvious use of the glottal sensors is in determining the presence of voicing. Theoretically, glottal sensors do not provide information about unvoiced speech but they can provide very accurate detection of voiced speech, including voiced consonants. Knowledge of this bit of information is invaluable for both parametric speech coding and speech enhancement. Of further importance, however, when coupled with information extracted from the noisy acoustic data, it also becomes possible to perform high-level segmentation of the speech.

2.2.2 Speech Segmentation

Accurate knowledge of high-level segmentation (voiced speech, unvoiced speech, and silence) is useful to any speech enhancement algorithm. Some enhancement algorithms yield more accurate noise estimates during silence, while other methods perform better during voiced speech. Further analysis may reveal other significant performance differences as a function of phonetic class. This observation suggests a potential advantage in running multiple enhancement algorithms in parallel to obtain multiple noise estimates, then applying low-level segmentation (vowel, fricative, plosive, etc.) in noise estimation, combining the various noise estimates optimally according to the performance of each, classified perhaps

by phonetic class or some other hierarchical representation. Low-level segmentation may also be useful in phonetic recognition.

Data from an acoustic microphone, P-mic at the throat location, and GEMS can be analyzed in tandem to obtain accurate estimates of the speech energy spectrum and periodicity. Accurate high-level segmentation, robust to a wide variety of audio and GEMS signals, is accomplished by means of heuristic methods using this data. These techniques can be extended to sub-bands to obtain improved accuracy. Low-level segmentation is accomplished by identifying subtler cues from both acoustic and non-acoustic sensors. For example, the characteristics of a noise-canceling microphone can be exploited by using an edge activity detector to mark the sudden spectral excursions that occur during plosives. The GEMS itself also provides cues to the presence of initial voiced plosives.

The segmentation algorithm incorporates energy thresholds of the acoustic and GEMS data, the autocorrelation sequence of the GEMS data, and added heuristics in its segmentation approach. These techniques are applied in a layered approach, refining the decision at every layer. This method is illustrated in Figure 1. The acoustic and GEMS signals are pre-processed to convert them to the desired format and to remove obvious noise components in the GEMS signal. The algorithm has been constructed to compensate for limited variations in the acoustic and GEMS signals. The heuristic is actually a state machine that tracks the transitions in the secondary segmentation result and uses the primary segmentation result to overcome errors that may have been introduced in the secondary stage. This algorithm is quite accurate and detects voiced onsets early. However, improving precision for unvoiced speech segments is still under investigation.

2.2.3 Noise Suppression

Algorithms utilizing these auxiliary sensors have been proposed in the past few years.A glottal correlation filter was introduced to enhance noisy speech using GEMS, an electromagnetic sensor that can detect the internal motions of the glottis. The filter is based on the correlation of speech signals and measured glottal excitation functions [5, 71, 72].

Pitch-Synchronous Comb Filtering: One of the major advantages of auxiliary sensors is their usefulness in detecting the exact glottal closure instants (GCIs). Using such detected GCIs, adjacent pitch periods of voiced speech are averaged together to obtain a slightly improved speech signal. But more significantly, as described above, an accurate estimate of the noise spectral shape during frames of voiced speech, i.e. high SNR (as opposed to noise-only frames), can be obtained for other processing.

Pitch-Synchronous Ephraim-Malah Suppression Rule (PS-EMSR): The Ephraim-Malah algorithm forms the basis for the MELPe noise enhancement module. This algorithm uses fixed frames. It is recognized, however, that if the framing were to be pitch-synchronous, many of the artifacts would be mitigated. Of course, this is only possible in noisy environments if the exact pitch epochs are detectable using non-acoustic sensors. Pitch-synchronous analysis provides an optimal solution to the signal uncertainty problem, which is addressed by non-optimal methods in MELPe. Moreover, the new algorithm exploits correlations between spectral bins, which are ignored by the EMSR used in MELPe. A vector form of the original EMSR is derived, which does not assume a diagonal covariance matrix, and the problem is solved in this framework. In some formulations, the solution does not have a closed form, in which case numerical techniques are applied.

Glottal Closure Detection (GCD): The GEMS signal provides a precise indication of glottal activity, although it may correspond more closely to tracheal wall motion or other physiological phenomena. Detecting the precise time of glottal closure is important for pitch-synchronous signal enhancement and coding. We have developed a highly robust algorithm that identifies those moments in time that are likely to be GCIs. The algorithm first determines the GEMS signals polarity and corrects it if necessary. Then it detects steepest descents and pitch in the GEMS data, and performs a maximum likelihood analysis to determine the most likely candidates for GCIs. A cross-correlation with the residual of the reverse-LPC filtered audio provides a rough estimate of the delay between the GEMS and acoustic signals, and the resulting pitch epoch information is then handed directly to the coder as well as to the segmentation and enhancement algorithms that require it.

LPC Estimation: Obtaining improved estimates of linear prediction coefficients (LPCs) of the audio improves the performance of the coder. The LPC estimation makes use of enhanced and noisy audio. GEMS is incorporated via the explicit use of the GCorr-enhanced speech discussed above. Our approach is to combine multiple outputs from the various noise suppression algorithms in a Multiple-Input Kalman filtering framework employing the Expectation-Maximization (EM) algorithm. The framework directly estimates the LPCs from the inputs to the Kalman filtering framework. The framework can be extended to a dynamic linear model, such as a Jump Markov Linear System (JMLS). A Multiple-Input Single-Output (MISO) Kalman filtering paradigm combines the outputs of the various speech enhancement algorithms, yielding LPC estimates in addition to enhanced speech. An autoregressive Hidden Markov Model (HMM) can be implemented that includes several states for clean and noisy speech vectors. The optimal set of states traversed is determined by Viterbi decoding. A state-dependent Kalman filter is then used to estimate clean speech and the LP parameters. Jump Markov Linear Systems combine HMMs and dynamic linear modeling. JMLS can be employed to further improve estimates of the LP parameters.

2.2.4 Speech Coding

It is clear that many of the issues that plague speech coding [29, 30, 31] in harsh environments would be mitigated should either an accurate characterization of the glottal flow pattern or a noise-robust low frequency sensor be available. Further, certain newly developed coders [27] are particularly sensitive to locating exact pitch epochs and as a result may not be robust to high noise conditions. The GEMS device can translate the gains made by these new coders to such environments through its ability to measure glottal activity.

2.3 Bandwidth Extension

Bandwidth extension (BWE) refers to methods that increase the frequency spectrum, or bandwidth, of electronic signals. Such frequency extension is desirable if at some point the frequency content of the signal has been reduced, as can happen, for example, during recording, transmission, or reproduction, mostly because of economical constraint. Most of the BWE methods heavily use signal processing - in fact, it is almost a premise that BWE

is a signal processing tool for achieving what is otherwise physically not possible.

BWE is a field that has been increasing attention in recent years. Although some work has been done in the early years of the twenty years, a much more systematic ad large-scale approach did not occur until recently. BWE for speech is the most mature area in this field, as the primary application (telephone speech) has existed for a long time. But as the widely application of Voice over Internet Protocol (VoIP) these years, BWE for speech has gained much more attention.

Bandwidth reduction implies a decrease in perceptual quality, and therefore BWE algorithms are employed as tools to enhance the perceived quality of reproduced sound. In most cases, BWE methods are post-processing algorithms, occurring just before sound reproduction, and the processing aims to compensate for limited bandwidth that is available in a prior part of the chain.

In recent years, there have been a significant number of publications on the bandwidth extension of speech signals [1, 28, 35, 38, 49, 68]. The BWE methods can be categorized based on the frequency range of interest. Some methods extend the low end of speech spectrum, other methods extend the high end of the spectrum. The classifications "low" and "high" in this sense are relative to the remaining speech spectrum, and should not be considered in the absolute sense. The second categorization is to realize "where" the signal bandwidth is actually extended: in the auditory system or in the waveform. In other words, psychoacoustic or physical. These four categories are indicated before:

1. *Low-frequency physical BWE:* The lowest frequency components of the signal are used to extend the lower end of the signal's spectrum. Such an algorithm can be used if the low-frequency bandwidth if the signal has been reduced in the storage or transmission; alternatively, the algorithm can be used for speech enhancement purposes, even if no prior bandwidth reduction had taken place. The loudspeaker will need to have an extended low-frequency response to reproduce the synthesized low frequencies.

2. *Low-frequency psychoacoustic BWE:* The lowest frequency components of the signal can not be reproduced by the loudspeaker, and are shifted to above the loudspeaker's

low cut-off frequency. This must be done in such a way as to preserve the correct pitch and loudness of the low-frequency.

3. *High-frequency physical BWE:* The highest frequency components of the signal are used to extend the higher end of the signal's spectrum. Such an algorithm can be used if the high-frequency bandwidth of the signal has been reduced in the storage or transmission.

4. *High-frequency psychoacoustic BWE:* The reproduction of high-frequency components is done in such a way as to preserve the correct pitch, timbre, and loudness.

In this thesis, we focus our work on the bandwidth extension algorithms that are generally aimed at recovering the spectral envelope for frequencies up to 8 kHz, given a speech signal with frequency contents below 4 kHz. This process removes the muffled sound quality, which is introduced when the speech signal is band limited to 4 kHz. In principle, the physical reconstruction of the acoustic bandwidth of speech signals can only be feasible if the algorithm has some *a* prior knowledge about the input signals. For example, if we consider an arbitrary signal that is sampled with a sampling rate of 8 KHz, and if there is no further information available on this kind of the signal components beyond the limit frequency of 4KHz. If, however, a mathematical model of the source of the signal is available, the situation is different: both the wideband signal as well as the bandlimited signal are determined by parameters of the common source model. Consequently exact knowledge of these source parameters would open up the possibility to reconstruct the complete wideband signal as it was originally produced. The parameters of the source, on the other hand, can be estimated from the characteristics of the bandlimited signal.

The typical architecture of a bandwidth extension system is shown in Figure 1. The whole algorithm can be viewed as two separate processes: the residual extension and the spectral envelope extension. An LPC analyzer extracts the spectral envelope from the input narrow-band signal. The residual extension module processes the resulting residual signal, while the envelope extension module predicts the wide-band spectral envelope, based on the narrow-band portion. The desired signal is then synthesized by using the wide-band

residual and the wide-band LPC coefficients.

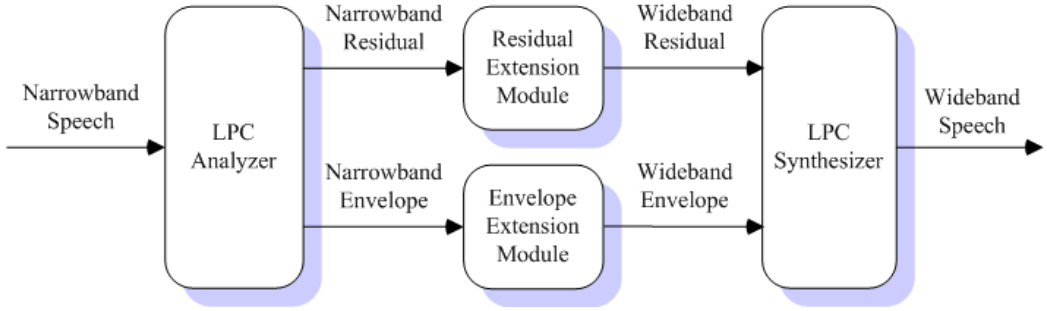


Figure 1: Block diagram of blind bandwidth extension

Residual Extension: The residual signal has a flat spectrum like white noise. In a voiced frame, such as vowels and semi-vowel consonants, the residual noise has periodicity. This appears as harmonic peaks in addition to the flat noise-like spectrum. These peaks occur in multiples of the pitch, the fundamental voice frequency of the speaker. Therefore, the task of the residual extension module is to increase the sampling rate, while keeping the whole spectrum flat. If there are harmonics in the narrow-band residual, the wide-band residual should also have the harmonic structure. There are two methods in common use to solve the problem.

1. **Nonlinear distortion method:** The narrow-band residual is first upsampled by interpolation and then fed into a nonlinear function. The distorted signal will have the desired bandwidth and harmonic structure over the whole spectrum. After the whitening filter, the spectrum is flattened and the wide-band residual is achieved [60, 89].

2. **Spectrum folding method:** The upsampling of the narrow-band residual is done by inserting zeros instead of interpolating. This is equivalent to folding the spectrum in the frequency domain. Since the low-frequency spectrum is flat and has harmonics, the resulting wide-band residual will also have a flat spectrum and harmonics in both the low-frequency part and the high-frequency part [60].

Envelope Extension The envelope extension is used to estimate the spectral shape

of the high-frequency components. It is the essential step in the bandwidth extension algorithms.

1. **Codebook mapping:** Codebook mapping is a popular method to achieve spectral envelope extension [7, 19, 79]. The codebook is generated from a large training database of speech. Eligible parameters to be used in the codebook are LPC coefficients, reflection coefficients, LSF, cepstral coefficients, etc. Besides the single codeword selection, an interpolation of several best selections was introduced in [10]. These approaches are memoryless since the estimation is based on the coming frame only. To reduce mismatching, when making the decision on the upcoming frame, the information from a number of previous frames is taken into consideration to find the codebook item with the highest probability. A codebook search method was proposed, based on hidden Markov models (HMM) [9, 47, 48, 75]. An enhanced version based on feature extraction is proposed in [50].

2. **Linear mapping:** Instead of codebook mapping, spectral envelope extension also can be done by linear estimation [11]. The set of parameters representing the narrow-band spectral envelope is first extracted from an input signal frame. Then, the corresponding vector representing the wide-band envelope is calculated by a group of linear filters.

2.4 Speech Evaluation

There are two principal perceptual criteria for measuring the performance of a speech processing algorithm. The quality of the enhanced signal measures its clarity, distorted nature, and the level of noise in that signal. The quality is a subjective measure that is indicative of the extent to which the listener is comfortable with the speech signal. The second criterion measures the intelligibility of the speech signal. This is an objective measure which provides the percentage of words that could be correctly identified by listeners. The words in this test need not be meaningful. The two performance measures are not correlated. A signal may be of good quality and poor intelligibility and vice versa.

2.4.1 Quality

Speech quality is a multi-dimensional term and its evaluation contains several problems [51, 61]. The evaluation procedure is usually done by subjective listening tests with response set of syllables, words, sentences, or with other questions. The test material is usually focused on consonants, because they are more problematic to synthesize than vowels. Especially nasalized consonants (/m/ /n/ /ng/) are usually considered the most problematic [8]. When using low bandwidth, such as telephone transmission, consonants with high frequency components (/f/ /th/ /s/) may sound very annoying. Some consonants (/d/ /g/ /k/) and consonant combinations (/dr/ /gl/ /gr/ /pr/ /spl/) are highly intelligible with natural speech, but very problematic with synthesized one. Especially final /k/ is found difficult to perceive. The other problematic combinations are for example /lb/, /rp/, /rt/, /rch/, and /rm/ [34].

Mean Opinion Score (MOS) is probably the most widely used and simplest method to evaluate speech quality in general. It is also suitable for overall evaluation of synthetic speech. MOS is a five level scale from bad (1) to excellent (5) and it is also know as ACR (Absolute Category Rating). The listener's task is simply to evaluate the tested speech with scale described in Table 1 below. In the same table a kind of opposite version of MOS scale, so called DMOS (Degradation MOS) or DCR (Degradation Category Rating), is presented. DMOS is an impairment grading scale to measure how the different disturbances in speech signal are perceived.

Table 1: Scales used in MOS and DMOS

	MOS (ACR)	DMOS (DCR)
5	Excellent	Inaudible
4	Good	Audible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Pair comparison methods are usually used to test system overall acceptance [53]. An average listener of a speech synthesizer will listen to artificial speech for perhaps hours per

day so the small and negligible errors may become very annoying because of their frequent occurrences. Some of this effect may be apparent if few sentences are frequently repeated in the test procedure [53]. Stimuli from each synthesizer are compared in pairs with all n(n-1) combinations, and if more than one test sentence (m) is used each version of a sentence is compared to all the other version of the same sentence. This leads total number of n(n-1)m comparison pairs. The category "equal" is not allowed [34].

One of the most commonly used objective measures to evaluate the quality of a speech enhancement system is the segmental SNR improvement. The gains were calculated as

$$SNR_{imp} = \frac{1}{L}\sum_{m=0}^{L-1} 10 \cdot \log_{10} \frac{\frac{1}{N}\sum_{k=0}^{N-1} n_m^2(k)}{\frac{1}{N}\sum_{k=0}^{N-1} [s_m(k) - \hat{s}_m(k)]^2} \tag{1}$$

where $s(i)$ is a speech signal, $n(i)$ is the background noise, L represents the number of frames in the voiced signal and N is the number of samples in the mth frame.

The log spectral distance (LSD) is computed in the Fourier domain. This is used instead of the computation of SNR in time domain, which will not produce meaningful results as the phase of the estimated high frequency components will not match that of the original 8kHz signal. The average log spectral distance, LSD, between the original speech, S, and the enhanced speech, X, is defined as:

$$LSD = \frac{1}{K}\sum_{k}^{k=K-1} \left[\frac{1}{\omega_s}\int_{-0.5\omega_s}^{0.5\omega_s} \left(\log_e |X(\omega)| - \log_e |S(\omega)|\right)^2 d\omega\right]^{0.5} \tag{2}$$

where K is the number of frames.

The Modified Bark Spectral Distortion (MBSD) measure [93] is an improved version of the Bark spectral distortion (BSD). It extends the BSD measure by incorporating the noise masking threshold into the algorithm to differentiate audible and inaudible distortions. The measure has been proven to be more correlated with speech quality than MOS.

2.4.2 Intelligibility

The intelligibility of speech refers to the accuracy with which a normal listener can understand a spoken word or phrase. Given the fact that some of the information communicated through speech is contained within contextual, visual and gestural cues, it is still possible to

understand meaning even if only a fraction of the discrete speech units are heard correctly. However, in large auditoria and places where reproduced speech is used, the listener has limited access to these cues and must rely more heavily upon the sound actually produced by the mouth.

Research into this area began with the development of telephone and telecommunication systems in the early part of this century. A product of this research was a quantitative measure for intelligibility based on articulation testing. This procedure (as described by Lochner and Burger [58]) normally consists of an announcer reading out lists of syllables, words or sentences to one or more listeners within the test enclosure. The percentage of these correctly recorded by the listeners is called the articulation score.

The science of articulation testing was substantially refined at Bell Telephone Laboratories and later at the Psycho-Acoustic Laboratory at Harvard University. From this later work, a set of phonetically balanced, mono-syllabic test lists were prepared, called the Harvard P.B.50 word score. In order to negate any influence of non-phonetic cues on the measured intelligibility, these word lists comprise only of meaningless or jumbled syllables. Thus, in order to be correctly recorded by the listener, each consonant and vowel sound must be clearly audible. As a further measure, many tests are conducted with the syllables embedded in a carrier phrase in an attempt to simulate fluent speech. There are now many derivations of this methodology (such as the Fairbanks rhythm method used by Bradley [3]), however, the resulting value is a percentage score of correctly recorded syllables. Thus the degree of intelligibility is considered to correlate with the average of these scores. This percentage becomes the measured speech intelligibility rating for that particular enclosure.

As stated before, normal connected speech can be understood even if some of the syllables are unintelligible. This is due to the fact that the listener can deduce the meaning from the context of the sentence. However, even under perfect conditions, the maximum word score normally attainable is about 95% due to unavoidable errors. A word score of 80% enables the audience to understand every sentence without due effort. In a room where the word score is closer to 70%, the listener has to concentrate to understand what is said whilst below 60% the intelligibility is quite poor.

There are several available methods of predicting Speech Intelligibility within an enclosure. The diagnostic rhyme test (DRT) [34, 59] uses monosyllabic English words that are constructed from a consonant-vowel-consonant sound sequence. In the DRT, one hundred and ninety two words are arranged in ninety-six rhyming pairs which differ only in their initial consonants. Listeners are shown a word pair, then asked to identify which word is presented by the talker. Carrier Sentences are not used. The DRT is based on a number of distinctive features of speech, and its test results reveal errors in discrimination of initial consonant sounds. The DRT is a quite widely used method. It provides lots of valuable diagnostic information how properly the initial consonant is recognized. The score is highly correlated to speech intelligibility and accepted as standard indication.

Most speech enhancement systems improve the quality of the signal at the expense of reducing its intelligibility. Listeners can usually extract more information from the noisy signal than from the enhanced signal by careful listening to that signal. This is obvious from the data processing theorem of information theory. Listeners, however, experience fatigue over extended listening sessions, a fact that results in reduced intelligibility of the noisy signal. Is such situations, the intelligibility of the enhanced signal may be higher than that of the noisy signal. Less effort would usually be required from the listener to decode portions of the enhanced signal that correspond to high signal to noise ratio segments of the noisy signal.

CHAPTER 3

SINGLE-CHANNEL NOISE SUPPRESSION USING BIOLOGICALLY INSPIRED TECHNIQUES

In single-acoustic-channel noisy environment, consider a speech signal $s(k)$ that is corrupted by statistically independent background noise $n(k)$, the noisy mixture $y(k)$ can be represented as

$$y(k) = s(k) + n(k) \tag{3}$$

When the ear is excited by a stimulus, the response patterns can be modeled as a bank of cochlear filters along the basilar membrane. Experimental measurements show a roughly logarithmic increase in the bandwidth of these filters, i.e, the filters are approximately constant-Q in their frequency response. The constant-Q cochlear filter bank, referred to as critical-band, thus provides a range of bandwidths to analyze the signal perceptually [76]. Apply a set of bandpass filters of critical bandwidths to $y(k)$. The m-th subband signal $y_m(k)$ is then obtained.

The objective of a subband-based noise suppression filter is to estimate the speech in the form

$$\hat{s}(k) = \sum_m h(m,k) \cdot y(m,k) \tag{4}$$

In popular used Wiener filter, $h(m,k)$ is constructed as

$$h(m,k) = \frac{\hat{\mathrm{S}}^2(\mathrm{m},\mathrm{k})}{\hat{\mathrm{S}}^2(\mathrm{m},\mathrm{k}) + \varsigma \cdot \hat{\mathrm{N}}^2(\mathrm{m},\mathrm{k})} \tag{5}$$

where $\hat{S}^2(m,k)$ and $\hat{N}^2(m,k)$ are the envelopes for subband speech and estimated noise, respectively. ς represents the level of noise over-substraction. Denote subband SNR $\varepsilon(m,k)$ as

$$\varepsilon(m,k) = \frac{\hat{S}(m,k)}{\hat{N}(m,k)} \tag{6}$$

Then, Equation (4) becomes

$$\hat{s}(k) = \sum_{m} \frac{\varepsilon^2(m,k)}{\varepsilon^2(m,k) + \varsigma} y(m,k) \tag{7}$$

The objective of the proposed algorithm is to use the biological correlates to estimate two parameters for the filter: one is the level of noise over-substraction, which will be derived from the probability of the presence of speech cues, and another is the level of boosting, which is applied to emphasize attenuated speech cues. Therefore, the proposed suppression filter is constructed as

$$\hat{s}(k) = \sum_{m} \frac{\nu(m,k) \cdot \varepsilon^2(m,k)}{\varepsilon^2(m,k) + \varsigma(m,k)} y(m,k) \tag{8}$$

$$\varsigma(m,k) = \frac{1}{\eta(m,k)} \tag{9}$$

where $\nu(m,k)$ is the level of boosting, which is indicated from the definition of temporal audibility saliency, and $\eta(m,k)$ represents the degree of conspicuousness of speech cues with the presence of background noise.

First, we define a soft-decision function for smooth estimation using a hyperbolic tangent form:

$$\begin{aligned} \Lambda(k) &= \frac{1 + tanh(k)}{2} \\ &= \frac{1}{2} \cdot (1 + \frac{e^{2k} - 1}{e^{2k} + 1}) \end{aligned} \tag{10}$$

The range of functional output is $(0, 1)$.

3.1 Frequency Importance Analysis

Employing the critical-band concept to separate the speech signal into subbands, which describes the frequency separation of the perceptual significance of acoustic stimuli, has been widely used for speech enhancement and coding [80]. In spectral substraction approaches [91], the frequency bins are divided into critical subbands and the speech signal envelopes are estimated separately in each band. But the dependency between each frequency bands has not yet exploited and used in the signal processing algorithms. In this section, the frequency importance analysis is used to derive a perception saliency that describes the

frequency-dependent attributes of speech cues in each frequency bands. The assumption is made that spectral variation in each critical subband plays a different role in the human perception of speech.

3.1.1 Mutual Information on Speech Source Classification

The predictability of the two observations can be stated as the mutual information between the two signals. The mutual information (MI, $I(x, y)$) between two variables x and y is described as

$$\begin{aligned} I(x,y) &= D[P(x,y) \| P(x), P(y)] && (11) \\ &= \int_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} dxdy && (12) \end{aligned}$$

where $P(x)$ and $P(y)$ are the densities of x and y respectively, and $P(x, y)$ is the joint density of x and y. D denotes the Kullback-Leibler divergence, also known as the relative entropy. The MI covers all kinds of linear and non-linear dependencies [15].

A MI analysis has been conducted for the relation of logarithm spectral energy in critical subbands and speaker-channel classification [92]. The mathematical relation is

$$I(c,m) = \sum_{c \in C} \sum_{e_m \in E} p(c, e_m) \log_2 (\frac{p(c, e_m)}{p(c)p(e_m)}) \quad (13)$$

where c and e_m are the source location and logarithm spectral energy in the m-th critical subband, respectively.

The function indicates the relevance of each subband to the discrimination of simultaneous sounds, the speech, and noise in noisy environments.

3.1.2 Spectral Relevance

In a noisy environment, the objective of speech detection is to give soft-decisions for the degree of speech presence based on the estimated subband SNRs. The detection of speech presence in noise is relevant to the classification of acoustic sources if we regard noise as an acoustic source to be identified. So, the above mutual information can be used as the frequency importance function for weighting the subband SNRs.

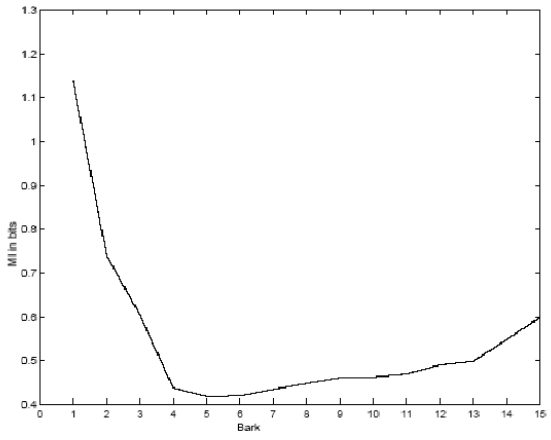


Figure 2: The MI between the speaker-channel labels and logarithm spectral energy in critical subband for speaker-channel classification.

Based on the results of Figure 2, the low- and high-frequency bands are most relevant for speech cue classification. A normalized weighting function, denoted as $w(m)$ for the m-th subband and expressed in Equation (30), describes the coherence of subbands correlated to the presence of speech cues.

$$\mathrm{w(m)} = \frac{2^{\mathrm{I(c,m)}}}{\sum\limits_{\mathrm{m}} 2^{\mathrm{I(c,m)}}} \tag{14}$$

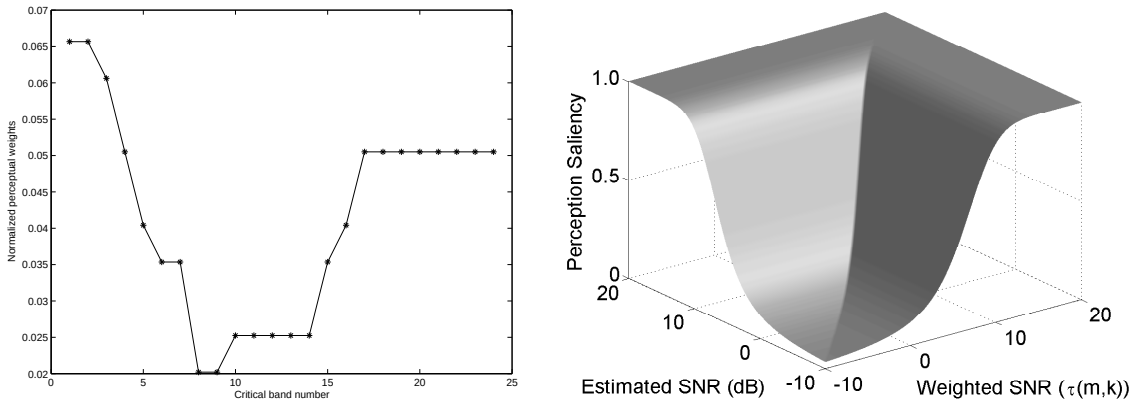
The results are shown in Figure 3. We define an estimation for the presence of speech cues in individual subbands, denoted as perception saliency. It is expressed as

$$\gamma(m,k) = \Lambda(\max\left[\epsilon(m,k), \gamma(k)\right] - 2) \tag{15}$$

The procedure generates frequency-dependent soft-decisions for determining speech presence.

3.2 Noise Estimation Based On Perceptual Detection

In practical applications of the speech processing systems, speech enhancement, speech coding, and speech recognition, the presence of background noise significantly deteriorates the performance. This is more true with the advent of mobile communication, in which the acoustic environments are varying and more complicated. Therefore, noise estimation is a very important component of the overall system, especially if the algorithm is required to handle nonstationary noise. The accuracy of noise estimation has a major impact on the



(a) (b)

Figure 3: Perception significance of critical band: (a) Frequency importance function in critical bands, (b) perception saliency function.

speech enhancement system. If the estimated level is too low, annoying residual noise will be introduced. If the level is too high, speech sounds will be muffled and intelligibility will be lost.

Current research aims at incorporating soft-decisions schemes to indicate the degree of presence of speech and to estimate noise during speech activity when possible. Such algorithms include minimum statistics (MS) [63], minima controlled recursive averaging (IMCRA) [12], Quantile based method [86], and soft-decision voice activity detector [85]. Time-frequency smoothing parameters and compensations are derived to track the nonstationary noise. The estimated noise variance may be too large, and occasionally attenuate low energy speech cues. The other drawback is that the estimation of speech presence probability is typically smoothed in frequency domain using only a small number of neighboring frequency bins.

In this thesis, we propose an efficient two-stage noise estimation algorithm. A recursive noise estimator is first applied to track the noise envelope based on a Gaussian model for the noise in critical sub-bands. The criterion is set to achieve a certain low degree of false alarm for the statistical model. At the second level, a perceptually inspired speech detector is employed. The coherence of the *a posterior* SNRs in all critical sub-bands is exploited and used to provide soft-decisions of the presence of speech cues.

3.2.1 Recursive Noise Estimator

When the noisy speech signals are received, the statistics of noise are unknown. In order to track the changing information quickly, in the first iteration of recursive noise estimation, the noise estimator therefore starts by assuming background noise level that is high, and a standard deviation that is large.

The critical-band filter bank does not provide a magnitude-phase representation of the signal as a DFT filter bank does. Therefore, the first step is to obtain a good estimate of the subband signal envelope for each subband. We apply a set of large frequency-dependent windows in smoothing the envelope of the noisy speech signals,

$$\bar{Y}^2(m,k) = \sum_{j=-J_2(k)}^{J_2(k)} \bar{b}(j) Y^2(m,k+j) \tag{16}$$

where the $\bar{b}(j)$ is a hamming window with the length of $[2J_2(m)+1]$, which is obtained based on the bandwidth of each subband.

$$\mathrm{J_2(m)} = min\{4 * \frac{\langle \frac{1.3}{\Delta\omega(\mathrm{m})} + 0.7273 \rangle - 1}{2}, L_{max}\} \tag{17}$$

The parameters $\Delta\omega(\mathrm{m})$ represents the bandwidth of m-th subband in radian, and $L_{max} = 0.05 * f_s$, which allows maximum window length of 50 msec in sample rate f_s. The operator $\langle * \rangle$ represents the rounding function in mathematics.

In this, the first, iteration, the noise envelope is estimated as

$$\bar{N}(m,k) = \min\{\bar{\beta} \cdot \bar{N}(m,k-1), \bar{Y}(m,k)\} \tag{18}$$

where $\bar{\beta}$ is set to be 1.0005, a large adapting number to track the fast changing nonstationary noise. A rough noise variance is then estimated as $\bar{\sigma^2}(m,k)$ adaptively.

Based on the obtained initial statistics information about noise , we model the subband noise as a Gaussian distributed variable with the mean of $\bar{N}(m,k)$, and the variance $\bar{\sigma^2}(m,k)$. The probability density functions (PDFs) are given by

$$f(N(m,k)|\bar{N}(m,k)) = \frac{1}{\sqrt{2\pi}\bar{\sigma}(m,k)} exp(-\frac{(N(m,k)-\bar{N}(m,k))^2}{2\bar{\sigma}^2(m,k)}) \tag{19}$$

In the second iteration, we need to get more accurate estimation of the noise variance and envelope. In this case, a set of smaller frequency-dependent smoothing windows are

applied to track the instantly rising envelope of the noisy signal more accurately.

$$\hat{Y}^2(m,k) = \sum_{j=-J_1(m)}^{J_1(m)} \hat{b}(j) Y^2(m,k+j) \tag{20}$$

where the hamming window length $[2 * J_1(m) + 1]$ is calculated as:

$$\mathrm{J_1(m)} = \frac{\langle \frac{1.3}{\Delta\omega(\mathrm{m})} + 0.7273 \rangle - 1}{2} \tag{21}$$

Based on the first iteration statistical information and noise envelope tracking, we propose a rough decision about speech presence:

$$v(m,k) = \begin{cases} 1, & if\ \bar{Y}(m,k) > \hat{\theta}_1(m,k) \\ 0, & otherwise \end{cases} \tag{22}$$

The threshold $\hat{\theta}_1(m,k)$ is expressed based on Gaussian analysis:

$$\hat{\theta}_1(m,k) = \bar{N}(m,k) + \hat{\rho}_1 \cdot \bar{\sigma}(m,k) \tag{23}$$

The parameter $\hat{\rho}_1$ is set to satisfy certain false alarm level ς_1

$$p(N(m,k) > \hat{\theta}_1(m,k) | \bar{N}(m,k)) < \varsigma_1 \tag{24}$$

Typically, $\varsigma_1 = 0.05$, and $\hat{\rho}_1 = 1.4$.

For each state of the signal, different strategy should be applied to track noise envelope. In noise-only regions, the noise estimator should move quickly and aggressively track signal envelope. Meanwhile, in the regions where speech presents, the adaptation of noise estimator need to be slow down in order to reduce speech distortion. Therefore, the adapting rates $\hat{\beta}(v(m,k))$ of noise estimator should be selective to each speech state. In this implementation, $\beta(1) = 1.0002$, and $\beta(0) = 1.00002$ are used respectively.

One important problem of the conventional noise estimation is the deterioration in speech onset regions, where speech signal is low-energy. This factor is crucial to speech intelligibility. Therefore, the noise estimator should be non-aggressive in these regions, in order to avoid the attenuation of the "significant" signal. We propose an alternative minimum tracking mechanism in these regions by a looking-ahead for the detection of strong speech. In the implementation, we use a $T = 50$ msec buffer of the noisy speech

signals. The minimum tracking indicator is activated, when strong speech is detected in the looking-ahead point, and the current signal is also identified to be speech presented. It is expressed in the following equation:

$$\psi_{\min}(m,k) = \begin{cases} 1, & if \bar{Y}(m,k) > \hat{\theta}_1(m,k) \\ & or \bar{Y}(m,k+T) > \hat{\theta}_2(m,k+T) \\ 0, & otherwise \end{cases} \tag{25}$$

where the threshold $\hat{\theta}_2(m,k)$ is defined as:

$$\hat{\theta}_2(m,k) = \bar{N}(m,k) + \hat{\rho}_2 \cdot \bar{\sigma}(m,k) \tag{26}$$

the parameter $\hat{\rho}_2$ is set to satisfy certain false alarm level ς_2, which should be much lower for this case.

$$p(N(m,k) > \hat{\theta}_2(m,k) | \bar{N}(m,k)) < \varsigma_2 \tag{27}$$

Typically, $\varsigma_2 = 0.005$, and $\hat{\rho}_1 = 2.3$.

Based on the above observations, the noise envelope tracked by the noise estimator is expressed as:

$$\begin{aligned} \hat{N}(m,k) = \quad & \min\{\hat{Y}(m,k), \psi_{\min}(m,k) \cdot \hat{N}(m,k-1) + \\ & (1 - \psi_{\min}(m,k)) \cdot \hat{\beta}(v(m,k)) \cdot \hat{\theta}_1(m,k)\} \end{aligned} \tag{28}$$

The noise variance estimation is also refined as $\hat{\sigma}(m,k)$.

3.2.2 Perceptually Inspired Speech Detector

In conventional noise estimation methods, the detection of speech is characterized by soft-decision speech presence probability. The detection is based on the statistical assumption that the real and imaginary part of a Fourier transform coefficients can be consider to be independent and can be modeled as zero mean Gaussian variables [4]. It was pointed out that this assumption holds only when speech stationary with a relatively small span of correlation and for a large frame size. The statistical based speech detection is therefore ineffective in practical applications.

In the proposed algorithm, we introduce a perceptually inspired speech detector exploiting the quantitative correlation of the acoustic stimuli and human hearing sensations. Since

the speech signals are divided using critical-band concept, which describes frequency separation of perceptually significance of acoustic stimuli, the coherence of all critical sub-bands is then analyzed in detecting a speech cue.

Based on the noise estimator in the first level, the *a posteriori* SNR in each sub-band at a time instant is defined as:

$$\gamma_1(m,k) = \frac{\hat{Y}(m,k)}{\hat{N}(m,k)} \tag{29}$$

The conventional noise estimation algorithm uses the correlation between a small number of frequency bins only. Actually, there exists certain correlation of all the critical sub-bands in the discrimination of different sounds [92]. In noise conditions, the problem is considered to be the discrimination of a speech cue and simultaneous acoustic disturbances. Therefore, the mutual information dependency $I(m)$ in critical sub-bands can be used for speech detection.

We defined a weighting function $w(m)$ as

$$\mathrm{w(m)} = \frac{2^{\mathrm{I(m)}}}{\sum\limits_{\mathrm{m}} 2^{\mathrm{I(m)}}} \tag{30}$$

The results are first used and illuminated in [40, 41].

The sub-band *a posteriori* SNRs is then weighted,

$$\gamma(k) = \sum_{m=1}^{K} w(m) \cdot \gamma_1(m,k) \tag{31}$$

We propose a speech presence probability function in mth sub-band at time instant k by explosive experiments,

$$\eta(m,k) = \frac{\tanh(1.2 * \max(\gamma_1(m,k), \gamma(k)) - 2) + 1}{2} \tag{32}$$

Bias compensation is shown effective in noise estimation [12], in our implementation, a time-varying frequency-dependent compensation parameter is adjusted by the speech presence probability.

$$\chi(m,k) = 1 + P(1 - \eta(m,k)) \tag{33}$$

Typically, $P = 0.7$.

The final noise estimation towards effective speech enhancement is then defined as:

$$\hat{N}(m,k) = \hat{Y}(m,k) \cdot \left\{ 1 - \frac{\hat{S}(m,k)}{\sqrt{\hat{S}^2(m,k) + \chi^2(m,k)\hat{N}^2(m,k)}} \right\} \tag{34}$$

where

$$\hat{S}(m,k) = \hat{Y}(m,k) - \chi(m,k)\hat{N}(m,k) \tag{35}$$

The overall algorithm requires much lower computational complexity compared to [63, 12]. The pseudocode is provided in Table 2.

Table 2: Pseudocode of the proposed noise estimation algorithm

- Divide $y(k)$ into critical sub-bands $y(m,k)$
- Apply frequent-dependent smoothing window $J_1(m)$, $J_2(m)$ to compute envelope $\hat{Y}(m,k)$ and $\bar{Y}(m,k)$
- Initialize noise variance $\bar{\sigma}(m,k)$ and $\bar{N}(m,k)$ in first 50ms
- For all time frame k
  - For all subband m
    (*Recursive noise estimator)
    · Compute $\bar{N}(m,k)$ using equation (4)
    · Update $\bar{\sigma}(m,k)$
    · Compute the high speech indicator $\psi_{\min}(m,k)$, the voiced indicator $v(m,k)$ in equation (11) and (8)
    ·Compute $\hat{N}(m,k)$ in equation (14)
    (*Perceptually inspired speech detector)
    · Compute perceptually inspired speech presence probability $\eta(m,k)$ in equation (18)
    ·Compute the time-varying frequency-dependent compensation $\chi(m,k)$ in equation (19)
    ·Update the final noise estimation $\hat{N}(m,k)$ in equation (21)

3.2.3 Performance Evaluation

The performance evaluation of the proposed noise estimation algorithm, and a comparison to the IMCRA method, consists of two parts. First, we test the tracking capacity of the noise estimators for nonstationary noise in low SNR condition. Second, we integrated the noise estimators into the speech enhancement systems, and measures the objective quality. The IMCRA method is integrated into a spectral subtraction based speech enhancement system. To show the advantage of the perceptually inspired speech detector, a modified

version of IMCRA, denoted as "M-IMCRA", was developed by incorporating the proposed speech presence probabilities. Finally, we implemented a speech enhancement system using the overall proposed noise estimation algorithm.

The clean signals used in our experiments are taken from TIMIT database. The signals are down-sampled to 8KHz and added with the noise signals from Noisex92 database. To track the performance in various conditions, we used two types stationary noise (car noise and room noise), and one type of nonstationary noise (babble noise).

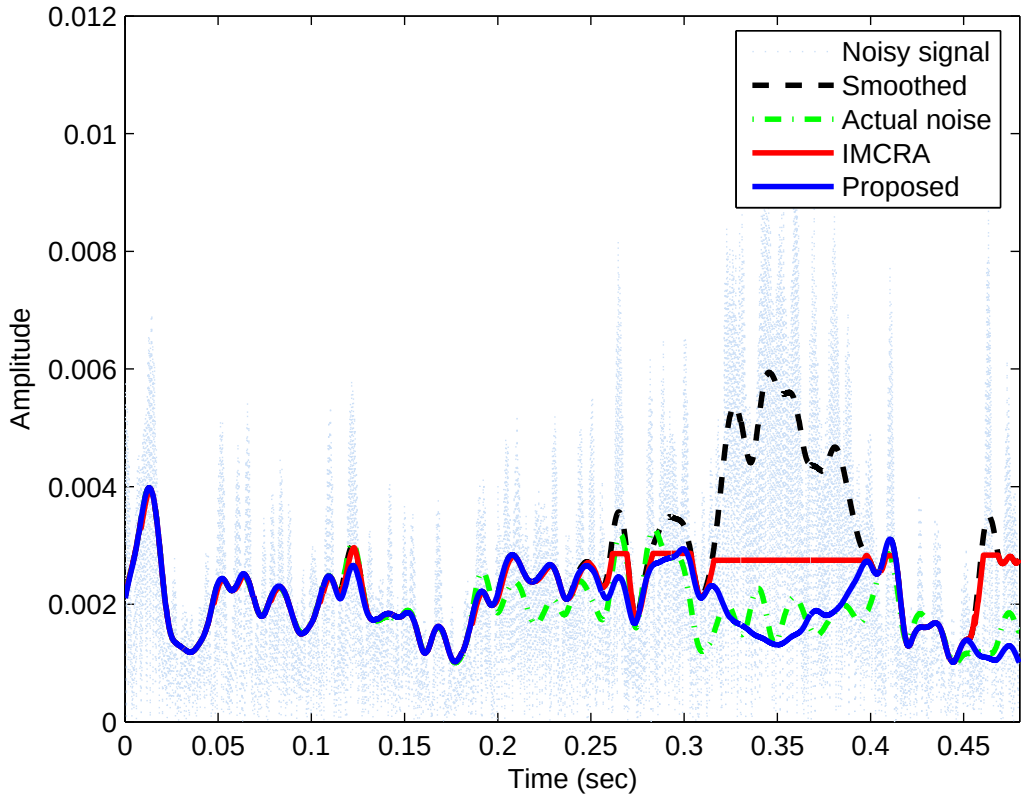


Figure 4: Example of noise envelope estimation using the proposed estimator and IMCRA. The speech signals (1700-2000Hz)are corrupted with additive babble noise(SNR=0.12dB).

The behavior of the proposed noise estimation algorithm is first illuminated in Figure 4 and Figure 5. In low SNR conditions, the proposed algorithm has much better ability to detect the presence of a speech cue, perceiving "interested" signal, and tracking non-stationary noise. Therefore, the proposed algorithm is especially effective in improving the segmental SNR in these sections. Although some degradation of segmental SNR in high

Table 3: Segmental SNR improvement in various noise conditions

Noise Type	Input Seg. SNR (dB)	Seg. SNR Improvement (dB)		
		IMCRA	M-IMCRA	Proposed
Car	-5	9.87	10.25	**10.67**
	0	8.42	8.64	**9.12**
	5	6.91	6.85	**7.17**
Room	-5	10.10	10.87	**11.16**
	0	8.69	8.93	**10.30**
	5	7.17	7.22	**7.53**
Babble	-5	6.15	5.99	**6.17**
	0	5.25	5.21	**5.36**
	5	4.09	4.26	**4.37**

speech sections. It will not affect the speech quality much, since the unnatural residual will be masked by the strong speech cues in spectrally neighboring. The improvement in low-energy phonemes and speech onsets will increase the intelligibility of speech, which is crucial in real-world application. It is also confirmed by informal listening tests.

Table 12 demonstrates the amount of noise estimation based on the segmental SNR improvement. The proposed system consistently performs the best in all candidates.

Modified Bark spectral distortion (MBSD) measure [93] is an improved version of the Bark spectral distortion (BSD). It extends BSD measure by incorporating noise masking threshold into the algorithm to differentiate audible and inaudible distortions. The measure has been proved to be more correlated with speech quality than Mean Opinion Score (MOS). To evaluate the objective quality, the MBSD improvement in various conditions are shown in Table 4. The proposed noise estimation system is consistently better than IMCRA. The "M-IMCRA", modified IMCRA featuring perceptually inspired speech detection techniques, in average, improves the MBSD performance too.

3.3 Biologically Inspired Suppression Rules

In the procedure of speech signal processing in the human auditory system, the stimuli are first preprocessed but still maintain their original character through the peripheral region. Then, the signals are led to the auditory sensation using neural processing in sensory cells.

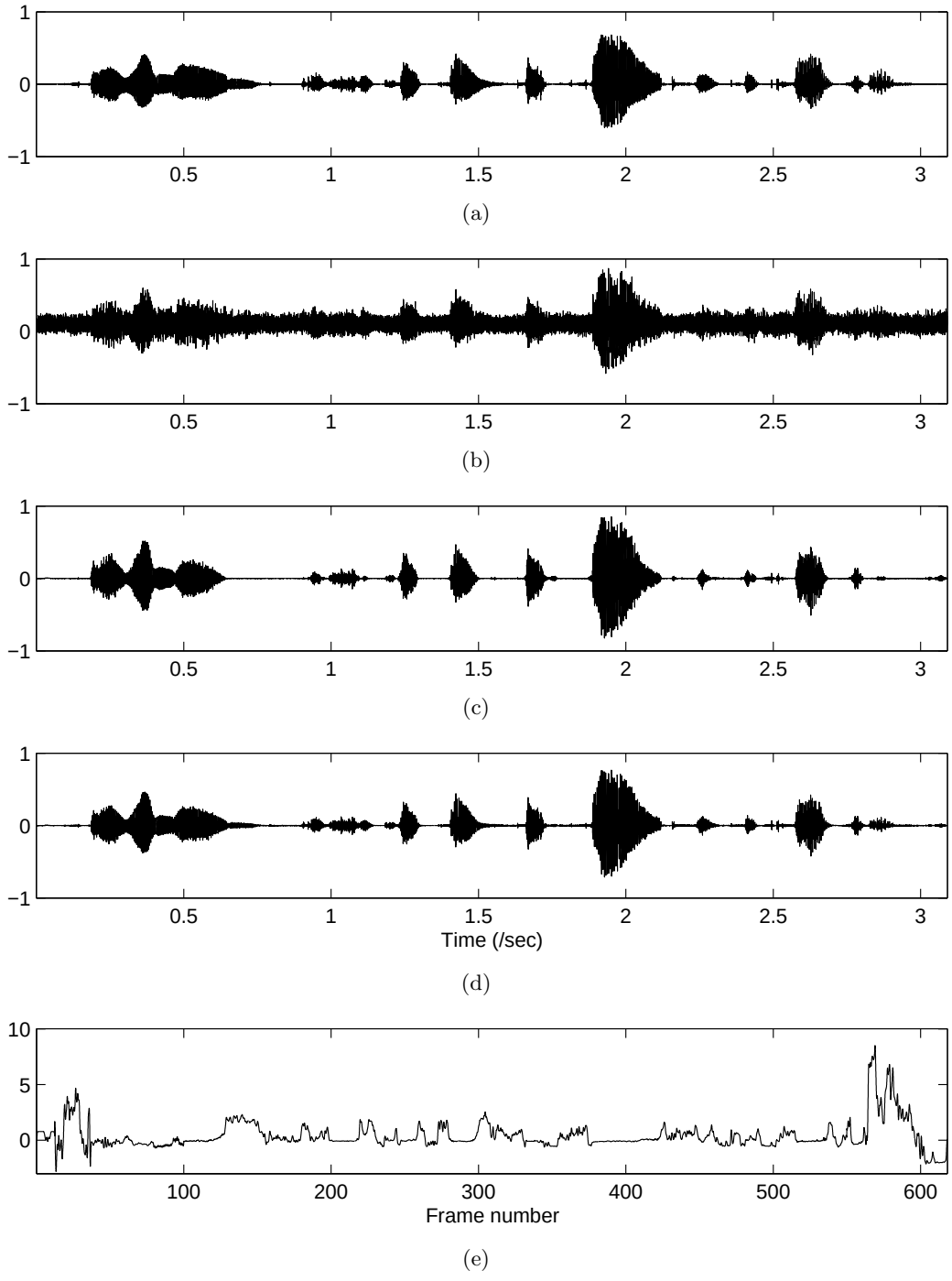


Figure 5: Example of speech enhancement using the proposed noise estimator and IMCRA. (a) Original clean speech, (b) noisy speech with additive babble noise at 0 dB segmental SNR, (c) enhanced speech using IMCRA, (d) enhanced speech using the proposed noise estimation algorithm, (e) trace of the gains of segmental SNR from the proposed over IMCRA.

Table 4: MBSD improvement in various noise conditions

Noise Type	Input MBSD (dB)	MBSD Improvement (dB)		
		IMCRA	M-IMCRA	Proposed
Car	5	4.09	4.31	**4.60**
	3	3.07	3.21	**3.52**
	1	1.41	1.50	**1.57**
Room	5	4.10	4.23	**4.77**
	3	2.69	2.79	**3.27**
	1	1.19	1.32	**1.53**
Babble	5	2.95	2.86	**2.97**
	3	1.64	1.71	**1.82**
	1	0.89	1.06	**1.13**

The perception of speech signals is highly correlated with biological correlates. Thus, a biologically inspired speech enhancement algorithm, that models human hearing mechanisms, can achieve perceptually improved performance. In the proposed method, a speech detector is derived to measure the degree of conspicuousness of "significant" speech signals with the presence of background noise [40, 41]. Three biological correlates are exploited, including perception saliency determining the speech cues by frequency sensitivity of the cochlear in the human auditory system, phoneme saliency indicating the attributes in the speech production mechanism, and audibility saliency illuminating psychoacoustic masking properties by the frequency-to-place transformation of the basilar membrane. The detector generates frequency-based soft-decisions that are used in determining the presence of speech cues and controlling the parameters of speech noise suppression. The block diagram is shown in Figure 6.

3.3.1 Phoneme Adaptation

The goal of phoneme saliency is to study the discriminatory features of each English phoneme class for the efforts on noise suppression.

In American English, speech sounds are classified into eight phonemes classes: vowels, semi–vowels, affricates, nasals, voiced/unvoiced plosives, and voiced/unvoiced fricatives. In the mechanism of speech production, speech sounds are determined by the source and vocal tract configuration. Each phoneme class has distinguishable acoustic properties. For

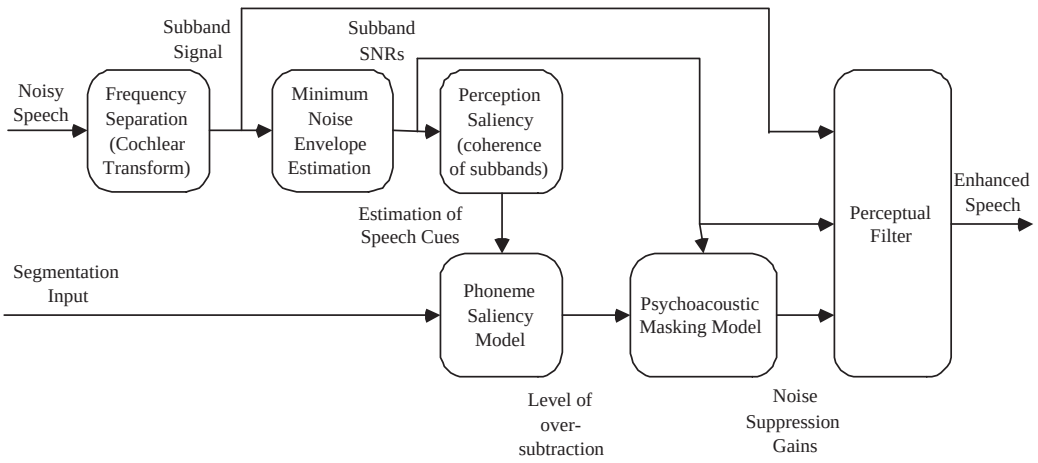


Figure 6: The block diagram of the proposed biologically inspired noise suppression (CQ) algorithm.

example, vowels have dominant frequency components at harmonic locations, especially at the formant locations. Unvoiced–plosives and fricatives have similar spectral distribution to that of white noise. Thus, the effects of background noise on these phonemes are also different.

It has long been known that each phoneme class has distinctive acoustic properties. The effects of background noise on the conspicuousness of "significant" signals are different for each class. Therefore, it is important to drive parameters selectively adjusted to speech signals based on phoneme content.

We classify speech sounds from the time–varying spectral characteristics, including the distribution and envelope. The significant components of fricatives are mainly distributed in the frequency bands above 1500Hz, the inverse of those of voiced plosives and nasals. The spectral contents of unvoiced-plosives, affricates, and whispers are almost flat in the whole frequency range, but with low envelope, thus, more susceptible to background noise. The others, such as vowels, have a high envelope and almost equal distribution. Regarding these distinguished acoustic characteristics and the susceptibility of ambient noise, we define enhancement-based classes of phonemes in Table 5.

In class 4 and 5, for example, aggressive noise estimation can be applied. But in other classes, since signals are relatively weak, they are more susceptible to background noise, especially in low SNRs. Noise estimation should be moderate in order to preserve more

Table 5: Enhancement-based classes of English phonemes.

Index κ	Included phoneme classes	Acoustic Properties
1	Fricatives	"Significant" signals concentrate in high bands
2	Voiced–plosives, Nasals	"Significant" signals concentrate in low bands
3	Unvoiced–plosives, Affricates	Weak "interested" signals exists in all bands
4	Vowels, Semivowels	Strong "interested" signals exists in all bands
5	Non–speech	No "interested" signals

interested components. A set of phoneme saliency functions is derived as:

$$\begin{aligned}\phi(m,k,1) &= \frac{1}{2}+\left\{\Lambda(20\cdot\log_{10}[\varepsilon(m,k)]-10)-\frac{1}{2}\right\}\cdot[1-\Lambda(2m-22)]+\Lambda(2m-22)\cdot \\ &\quad \left\{\Lambda(20\cdot\log_{10}[\varepsilon(m,k)])-\frac{1}{2}\right\} \qquad (36)\\ \phi(m,k,2) &= \frac{1}{2}+\left\{\Lambda(20\cdot\log_{10}[\varepsilon(m,k)]-10)-\frac{1}{2}\right\}\cdot\Lambda(2m-20)+[1-\Lambda(2m-20)]\cdot \\ &\quad \left\{\Lambda(20\cdot\log_{10}[\varepsilon(m,k)])-\frac{1}{2}\right\} \qquad (37)\\ \phi(m,k,3) &= 1 \qquad (38)\\ \phi(m,k,4) &= 0.765 \qquad (39)\\ \phi(m,k,5) &= 0 \qquad (40)\end{aligned}$$

The definition indicates the probability of the existence of speech cue with a value ranged in $(0,1)$, as shown in Figure 7.

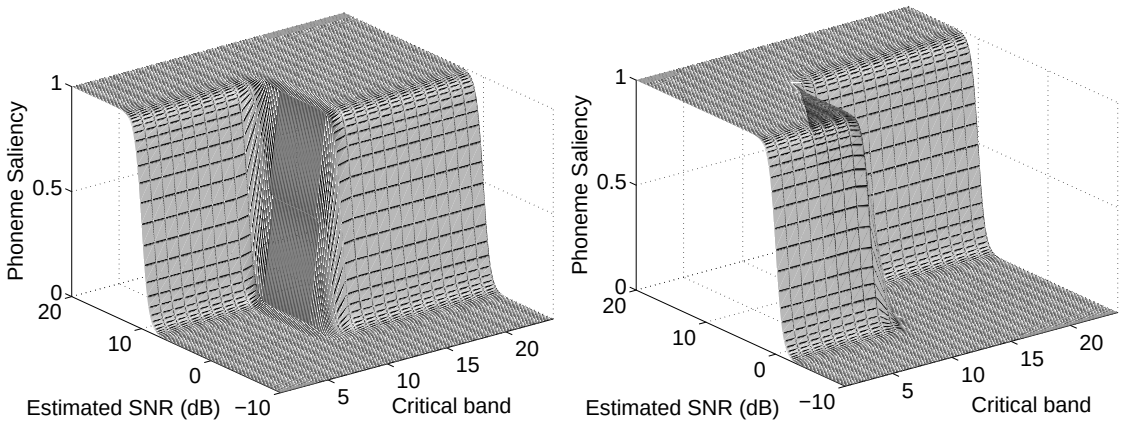
A function represents the combination of perception saliency and phoneme saliency gives a refined noise over–substraction parameter:

$$\mu(m,k) = 1+[1-\gamma(m,k)]\cdot\{\alpha_{\min}+(\alpha_{\max}-\alpha_{\min})\cdot(1-\rho(m,k,\kappa))\} \qquad (41)$$

where $\alpha_{\max}$ and $\alpha_{\min}$ are the maximum and minimum value of noise over–substraction. In the proposed method, we set $\alpha_{\max}=3$ and $\alpha_{\min}=1.3$.

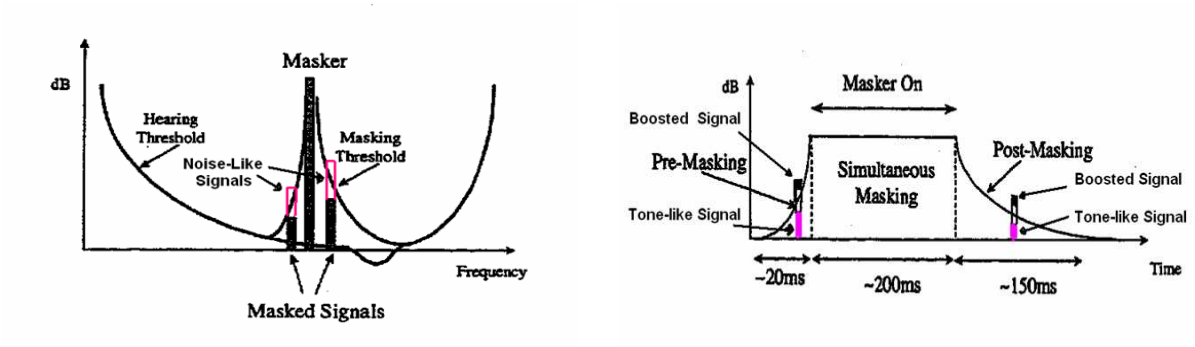
3.3.2 Psychoacoustic Masking Model

Psychoacoustic masking effects are the occlusions of one sound by another loud sound. The basic idea of the psychoacoustic masking model in speech enhancement is to force the



(a) (b)

Figure 7: Phoneme saliency: (a) fricative (b) voiced plosive and nasal.



(a) (b)

Figure 8: Basic idea of psychoacoustic masking (PAM) model in speech enhancement, (a) spectral masking, (b) temporal masking.

undesired residual artifacts into an inaudible level and retain the audibility of the weak interested signals from the derived noise masking thresholds, as indicated in Figure 8

Spectral masking, also referred to as simultaneous masking, takes place among the different frequency components of sounds occurring at the same time. A masking threshold appears below the components gathered within one critical band. The masking effects depend on the signal energy, the frequency of the masker, and the tonalities of the masker and the maskee.

The steps for calculating the noise masking threshold $T_m(k)$ from the signal envelopes in each subband are as follows:

1. Convolution with a spreading function to take into account masking between different

critical bands, as shown in Figure 9(a).

2. Subtraction of a relative threshold offset depending on the noise-like or tone-like nature of the masker. The simplified thredshold offset is based on the fact that the speech signal has a tone-like nature in lower bands and a noise-like nature in higher bands [83], as shown in Figure 9(b).

3. Renormalization and comparison with the absolute threshold of hearing.

We define audibility saliency $\theta(m,k)$ to indicate the probability of residual musical artifacts in a subband signal. The parameter is calculated as

$$\theta(m,k) = \Lambda(\frac{A_s}{T_{\max}(k) - T_{\min}(k)} \cdot [T(m,k) - T_M(k)]) \tag{42}$$

where A_s equals 10 derived for the smoothing function, and

$$T_M(k) = \frac{T_{\max}(k) + T_{\min}(k)}{2} \tag{43}$$

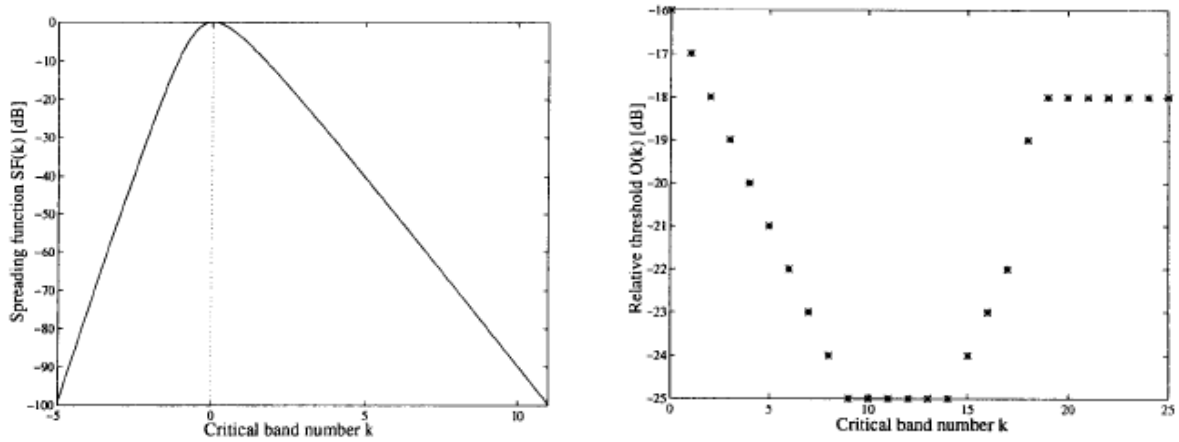
An assumption is made for the following consideration: if the masking threshold is high, residual noise will be naturally masked and inaudible. However, if the masking threshold is low, residual noise will be annoying to the human listener and it is necessary to reduce it. Therefore, Speech saliency $\eta_m(k)$ is defined to measure the presence of speech cues and residual artifacts.

$$\eta(m,k) = \frac{1}{\mu(m,k) + [1 - \theta(m,k)] \cdot [\mu_{\max}(k) - \mu(m,k)]} \tag{44}$$

Two main temporal masking phenomena [37] have been observed in the human audition: pre-masking and post-masking.

In low SNR cases, the onset regions of speech section are highly degraded by ambient noise. It is important for the intelligibility of start consonants. Thus, a backoff technique is proposed for noise envelope estimation using the pre–masking thresholds from posterior SNR. The procedure follows these steps:

1. Perform the first level noise envelope estimation.



(a) (b)

Figure 9: Spectral masking characteristic functions: (a) spreading function (b) simplified relative threshold offset

2. Derive the temporal pre-masking threshold curve when a strong speech section is detected.

3. Determine the onset of speech by finding within pre–masking region of each subband the first place where the residual signal envelope exceeds the masking threshold.

4. Move noise estimation mechanism back-off to the onset point to slow down adaptation rate for the effort of perceiving more interested signal .

To retain the significance of speech signals, both pre–masking and post–masking are used to post–process the enhanced signals after noise estimation.

Assume that within temporal masking regions of the original noisy mixture, the signal envelope ($S(m,k)$), masking threshold ($T(m,k)$), and audibility level ($L(m,k) = S(m,k) - T(m,k)$) are provided. The corresponding parameters for the enhanced speech signals are $S'(m,k)$, $T'(m,k)$, and $L'(m,k) = S'(m,k) - T'(m,k)$. The desired parameters for the post-processed signals are $\hat{S}(m,k)$, $T'(m,k)$, and $\hat{L}(m,k) = \hat{S}(m,k) - T'(m,k)$.

The range of audibility level [$L_{min}(m,k)$-$L_{max}(m,k)$] is

$$L_{\max}(m,k) = S(m,k) - T'(m,k) \quad (45)$$

$$L_{\min}(m,k) = L'(m,k) \quad (46)$$

An interpolated decision for audibility level goes to

$$\hat{L}(m,k) = L_{\min}(m,k) + \frac{1}{2}\left[L_{\max}(m,k) - L_{\min}(m,k)\right] \cdot \left[\gamma(m,k) \cdot \phi(m,k) + \theta(m,k)\right] \quad (47)$$

The temporal audibility saliency indicates the level of boosting for retaining the same audibility level of significant signals within temporal masking. The definition is represented as

$$\nu(m,k) = \frac{\hat{L}(m,k)}{L'(m,k)} \quad (48)$$

For signal outside temporal masking, $\nu(m,k)$ is set to 1.

3.4 Performance Evaluation

3.4.1 Objective Measure

Objective measures for this biologically inspired noise suppression algorithm were performed on TIMIT database. Sample speech spectrograms are provided in Figure 10.

The Modified Bark Spectral Distortion (MBSD) measure [93] is an improved version of the Bark spectral distortion (BSD). It extends the BSD measure by incorporating the noise masking threshold into the algorithm to differentiate audible and inaudible distortions. The measure has been proven to be more correlated with speech quality than Mean Opinion Score (MOS).

The segmental SNR improvement of speech is an important measure for demonstrating the amount of noise suppression. The gains were calculated as

$$SNR_{imp} = \frac{1}{L}\sum_{m=0}^{L-1} 10 \cdot \log_{10} \frac{\frac{1}{N}\sum_{k=0}^{N-1} n_m^2(k)}{\frac{1}{N}\sum_{k=0}^{N-1}\left[s_m(k) - \hat{s}_m(k)\right]^2} \quad (49)$$

where L represents the number of frames in the voiced signal and N is the number of samples in the mth frame. Initial results on MBSD and segmental SNR improvement are shown in Figure 11.

3.5 Aurora 2 Noisy Speech Recognition

Recently the design of automatic speech recognition (ASR) systems for use in personal and mobile electronic devices has been seeing a tremendous growth. The design of robust

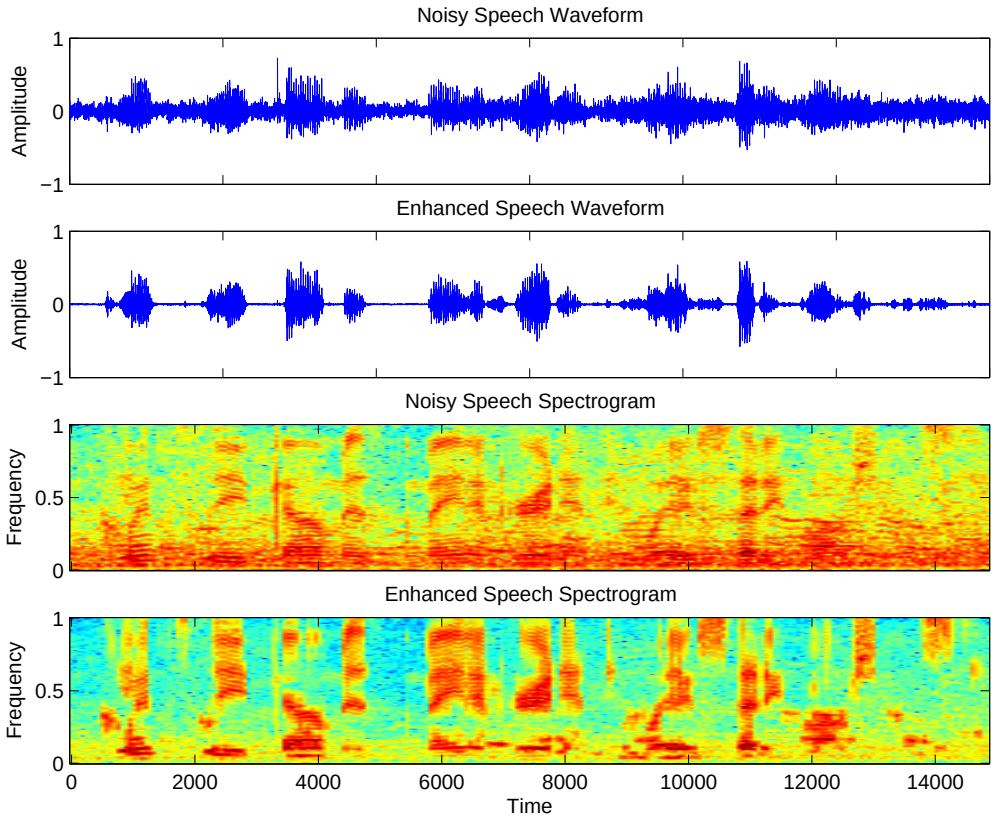


Figure 10: Waveforms and spectrograms of noisy source (degraded with babble noise, SNR=5dB) and enhanced speech signal of the proposed algorithm.

ASR systems for use in mobile environments poses several research challengers. First, these systems must perform without degradation in a variety of environmental conditions, where the input speech is corrupted by background noise. Second, the implementation of these systems is constrained by the limited resources available in wireless devices. In a distributed speech recognition (DSR) environment, features are extracted from the speech signal at the remote location and the recognition is performed in a centralized server. One is the solutions to the problem of designing robust ASR systems is to employ noise suppression algorithms prior to the feature extraction by the DSR system. Alternatively, the recognition system can be trained so that the speech models match the noisy environment. While the former requires the incorporation of a suitable noise suppression algorithm in the front-end (feature extraction process), the later approach is related to the modification of back-end (the speech models that perform the recognition task).

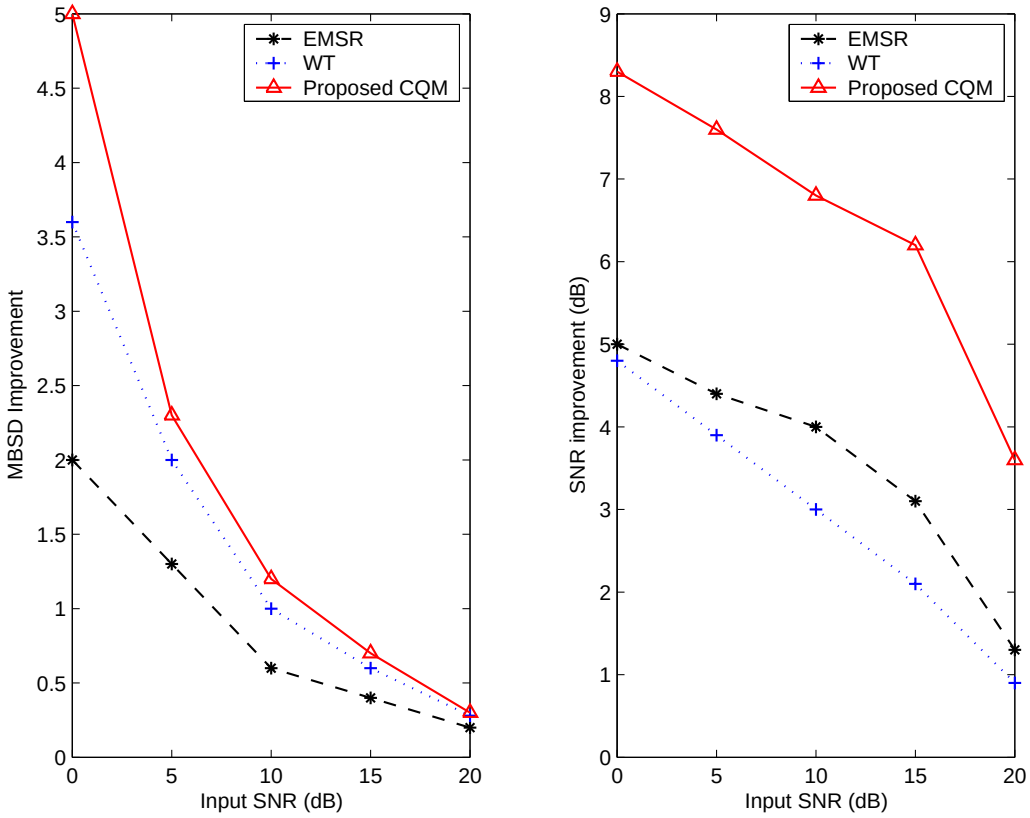


Figure 11: Comparative performance, in terms of MBSD and segmental SNR improvement measures for 50 TIMIT sentences corrupted by babble noise at 0–20dB SNR.

Given the need to have a common platform where researchers could test their noise pre-processing and ASR algorithms, and compare their results fairly, the Aurora DSR Working Group defined a set of connected digit string recognition experiments called the Aurora-2 task [77]. The basic Aurora-2 task consists of a standard front-end to extract the feature vectors and a standard back-end to perform the connected digit string speech recognition. Also a speech database was provided and an evaluation criterion was defined. This common platform is commonly referred as Aurora task. The Aurora-2 task provides a speech corpus referred as Aurora-2.0, which is a down-sampled subset of the TI Digits corpus. This database was artificially corrupted using different kinds of noise, including subway, babble, car, exhibition hall, restaurant, street, airport, and train station noise.

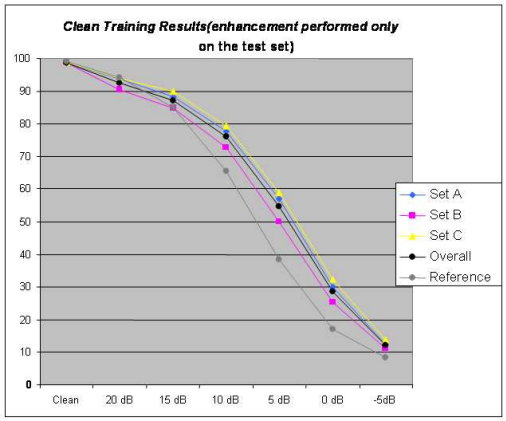
The Aurora-2 task defines two training modes: (a) clean training mode in which the recognition engine is trained on clean data alone and (b) multi-conditional training where

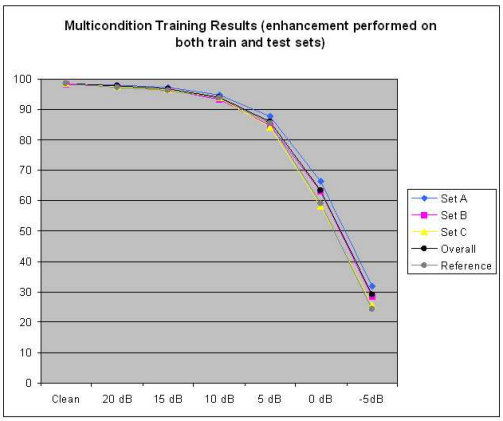
training is done using both clean and noisy data. The clean training database is composed of 8440 digit strings from TIDigits [61] that have been filtered by the G.712 characteristic filter without any addition of noise. The multi-conditional set consists of the same data as the clean set, but the data are divided into 20 subsets, each with 422 utterances. These 20 subsets represent 4 different noise conditions (suburban train, babble, car and exhibition hall) at 5 different SNR levels. The files are first filtered by the G.712 [46] filter prior to noise addition. Three testing sets are provided for the evaluation of the Aurora-2 task. Each set has 4 subsets of 1001 utterances obtained from the TI Digits test database. The first testing set is set A that contains four sets of 1001 sentences, corrupted by subway, babble, car, and exhibition hall noises, respectively, at different SNR levels. Thus, the noise types included in this set are the same as those in the multi-conditional training. The second set, set B contains 4 sets of 1001 sentences each, corrupted by restaurant, street, airport, and train station noises at different SNR levels. These noise types are different from the ones used in the multi-conditional training. The test set C contains 2 sets of 1001 sentences, corrupted by subway, and street and airport noises. The data set C was filtered with the MIRS filter [46] before the addition of noise in order to evaluate the robustness of the algorithm under convolutional distortion mismatch.

A version of CQ without segmentation input was used in these tests. The results (Figure 12) indicate the improvement of word accuracy for a machine as a comparison of the proposed algorithm compared with state-of-the-art algorithms.

3.6 Summary

The proposed algorithm conducts experimental trials for audio noise suppression modeling speech the perception mechanism of the human auditory system. Perceptually criterion and phoneme adaptive mechanism are imposed on the noise suppressor. The amount of suppression of this multiple–state algorithm is a non-linear function of the detected saliency and corresponding phoneme class. The algorithm effectively removes background noise. Although contained some residual noise, the enhanced outputs suffer less spectral distortion. Significant improvement on both speech quality and intelligibility were achieved. The





Performance relative to Baseline				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	7,91%	5,51%	13,60%	8,30%
Clean Only	20,73%	20,12%	14,09%	19,33%
Average	14,32%	12,82%	13,85%	**13,82%**

Figure 12: Speech enhancement performance for speech recognition view in Aurora-2 database

potential of more refined phoneme adaptive strategies and robust functions to segmentation errors can be further worked on. And better results can certainly be obtained.

An improved perceptually inspired speech enhancement algorithm using a psychoacoustic model is also introduced in this chapter. The objective is to limit speech distortion and improve speech intelligibility by controlling the enhancement gains by the biological correlates. A psychoacoustic model is proposed exploiting spectral saliency, phoneme adaptation and masking properties. The improvement is indicated from objective measures. Also, informal listening test conveys improved speech quality.

CHAPTER 4

MULTI-SENSOR NOISE SUPPRESSION FOR HARSH ENVIRONMENTS

This chapter describes a speech enhancement system that significantly improves speech intelligibility of noisy speech in the context of a speech coder in low SNR conditions. The system uses two state-of-the-art non-acoustic sensors, a general electromagnetic motion sensor (GEMS) that detects the internal motions of glottis, and a physiological microphone (P-mic) that measures vibrations of the skin associated with speech. Both sensors are relatively immune to ambient acoustic noise, but provide incomplete information of speech. A high resolution of speech segmentation is developed using the side information from these sensors. In the proposed system, referred to as "CQ–GCORR", the strengths of two algorithms, a perceptually motivated constant-Q (CQ) algorithm and an enhanced glottal correlation (GCORR) algorithm, are combined. The CQ algorithm employs a perceptually inspired signal detection technique to estimate the presence of speech cues in low SNR conditions. To reduce annoying artifacts, a state-dependent mechanism discriminating the distinct acoustic properties of each phoneme, and a psychoacoustic masking model are used to control enhancement gains. The GCORR algorithm extracts the desired speech signal from the noisy mixture, using a speech-GEMS correlation estimation of the speech signal with the glottal waveform supplied by GEMS. Both subjective and objective experiments were performed for low-bit-rate speech coding in a variety of noise conditions. The results indicate the improvements on both speech intelligibility and quality.

4.1 Secondary Sensors

It is known that the speech acquired by normal acoustic microphone is easily corrupted by the ambient noise. Today's technology has opened whole new realms in transducer technology. Ear microphones, Tooth microphones, Bone Conduction transducers, Electrical

and Micro-Radar based transducers are but a few of the devices available to the designers of communication systems. Each brings it's own benefits and drawbacks. The signals measured from these non-air conductive sensors exhibit significant attenuations on ambient noise, but provide incomplete information about the speech signal. The acquired side information can benefit the effort of speech enhancement, especially in low signal-to-noise ratio (SNR) cases.

4.1.1 Glottal Electromagnetic Sensor(GEMS)

Newly developed electromagnetic (EM) near field sensors (refer to the technology transfer web site) provide a capability for measuring EM wave reflections from speech organ interfaces in a non invasive, safe, fast, portable, and low cost fashion. These devices have similarities to some far-field radar systems except that their power levels are very low, their measurements are commonly in the near-field, and their rate of data acquisition (e.g., prf) is very high and very flexible. They are being used in investigations for other applications such as heart function and mechanical vibration sensing. In particular, they make possible the real-time measurements of the positions and motions of the human vocal articulators during speech production. The measurements to date include the motions of the glottis (i.e., vocal folds), lips, tongue, jaw, and velum. Examples of vocal fold measurements (in real time) gives the pitch period, enable noise removal, and enables pitch synchronous deconvolving of the corresponding excitation from the acoustic output to obtain improved quality speech transfer functions. Similarly, EM sensor measurements of jaw motion with acoustic speech provide constraints on the sound being articulated for speech recognition, and can be used for "talking head" and video image synchronization.

4.1.2 Physiological Microphone (P-mic)

The Army Research Laboratory (ARL) has developed a new method of measuring human physiological stress parameters. The so-called physiological microphone (P-mic) consists of an acoustic sensor positioned inside a fluid-filled bladder in contact with the body [82]. Packaging the sensor in this manner minimizes outside environmental interferences, and

signals within the body are transmitted to the bladder with minimal losses. This fluid-coupling technology comfortably conforms to the human body, and enhances the signal-to-noise ratio (SNR) of human physiology to that of ambient noise. An acoustic sensor system can detect changes in a person's physiological status resulting from exertion or injuries such as trauma, penetrating wound, hypothermia, dehydration, heat stress and many other conditions or illnesses. A sensor contacting the torso, head, or throat region picks up the wearer's voice very well through the flesh, with fidelity sufficient to be used as a hands-free voice activation mechanism. The P-Mic is worn like a collar, and has a silicon contact sensor which is placed slightly to the left or right of the throat, due to the symmetrical nature of the throat. The P-Mic is small and lightweight.

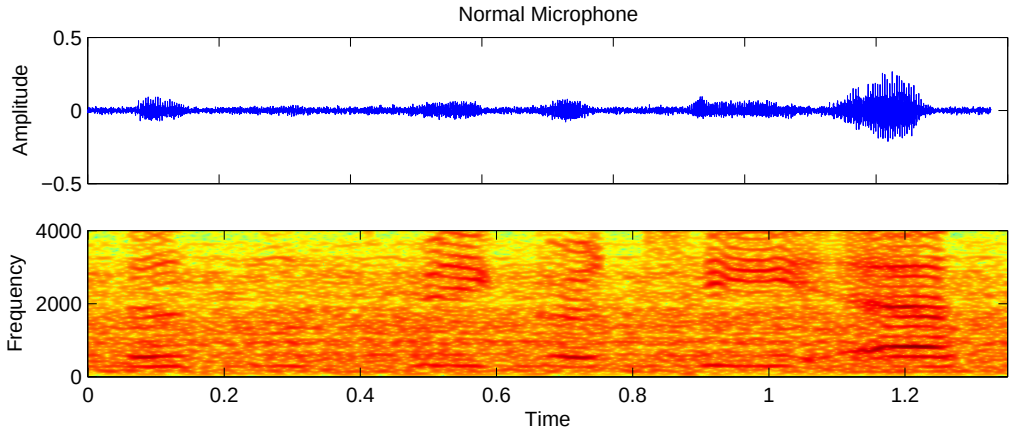
4.2 Information Analysis

In order to be effective, the measurements from the secondary sensors must be related to the underlying clean speech signal. Stated otherwise, knowledge of the outputs of the sensors must reduce the uncertainty in our knowledge of the speech signal. The predictability of the speech signal from the sensor measurement can be stated as the mutual information between the two signals.

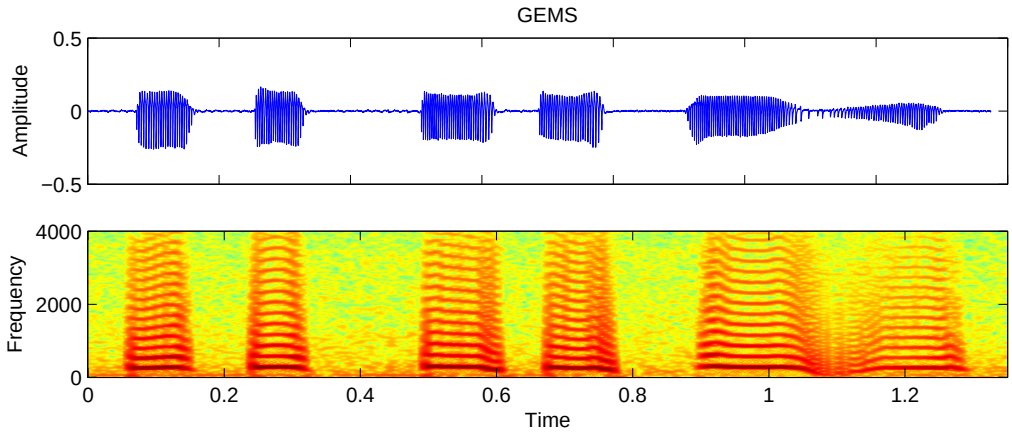
In case the statistical distributions of the variables are unknown and only a limited amount of samples of the variables are available for measurement, a non-parametric estimator is proposed in [16]. The algorithm approximates the mutual information arbitrarily closely in probability by calculating relative frequencies on appropriate partitions of the data space and achieving conditional independence on the rectangles of which the partitions are made.

Reviewing the objectives of employing a secondary sensor in robust speech processing , the qualification of a secondary sensor in robust speech processing can be summarized in information theoretic terms as follows:

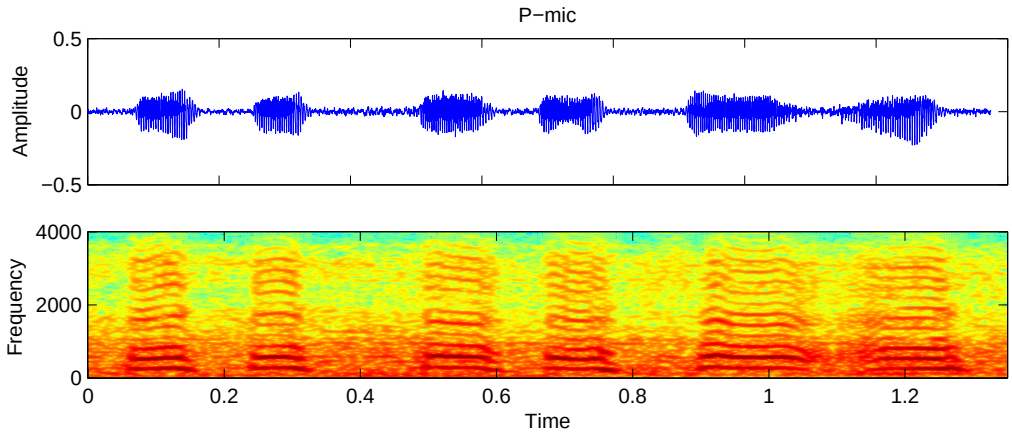
- High dependency between the outputs of the secondary sensor X and clean speech S, i.e. $I(X, S)$ is large.
- High independence of the outputs of a secondary sensor X and noise N,i.e. $I(X, N)$



(a)



(b)



(c)

Figure 13: Sample waveforms and spectrograms of the measured signals from (a) normal microphone; (b) GEMS; (c) P–mic.

is low.

In recordings obtained from high-noise environments, the second condition may also be stated as a requirement of low $I(X,Y)$, i.e. of independence between the doppler and noisy speech measurements. Given these criteria, the robustness of a secondary sensor can be represented as the normalized change of mutual information in noisy environments.

$$\Delta I(X,Y\|SNR) = \frac{I(X,S) - I(X,Y\|SNR)}{I(X,S)} \tag{50}$$

The greater the value of $\Delta I(X,Y\|SNR)$ the more useful the measurements of the sensor can be expected to be in processing highly noisy speech.

The MI analysis of recordings from GEMS, P-mic and EGG sensors is listed in Table 6. The results confirm the observations in [42, 66] that GEMS contains more secondary information about speech than P-mics and EGG, and is also more robust than the others two. As described above, P-mic recordings contain some level of acoustic noise. All of these sensors have been applied to robust speech processing and have produced improved performance in voice activity detection and speech enhancement [66].

Table 6: Mutual Information between the sensor outputs and acoustic signals

Clean Environment	GEMS	P-mic	EGG
I(X,S)	0.272	0.075	0.091

Noise	Office	Tank	Shoot	Helicopter
$\Delta I(X,Y)$	(23dB)	(1dB)	(13dB)	(3dB)
GEMS	0.202	0.993	0.743	0.996
P-Mic	0.280	0.693	0.027	0.640
EGG	0.044	0.912	0.626	0.967

4.3 Glottal Correlation Filter

It is well-known that, in speech production mechanisms, the acoustic speech signal is generated by the frequency shaping of the glottal excitation signal by the vocal tract. Those internal motions of glottis, which are measured by the GEMS device, are independent of ambient noise. The idea of the glottal correlation filter is to extract acoustic speech signal, that

statistically correlates with glottal excitation, from noisy mixture [42]. The mathematical descriptions are listed below.

4.3.1 Glottal Correlation Property

Table 7 lists the variables in frequency domain :

Table 7: Definition of Variables used in GCORR

$X(f)$: Signal acquired by acoustic microphone
$\bar{G}(f)$: Signal acquired by GEMS
$S(f)$: Clean Speech
$N(f)$: Noise signal acquired by acoustic microphone
$T(f)$: Vocal tract transfer function
$G(f)$: Glottal excitation signal
$H(f)$: Transfer function between GEMS signal and glottal excitation
$M(f)$: Noise signal acquired by GEMS

The frequency domain signals, $X(f)$, $S(f)$ and $\bar{G}(f)$, can be described as:

$$X(f) = S(f) + N(f) \tag{51}$$

$$S(f) = T(f)G(f) \tag{52}$$

$$\bar{G}(f) = H(f)G(f) + M(f) \tag{53}$$

The real and imaginary part of a Fourier transform coefficient background noise ($N(f)$) can be considered to be independent and can be modeled as zero mean Gaussian random variable [4]. The glottal excitation is a human internal glottis movement, the source of clean speech, it is uncorrelated to the external disturbance. Therefor, the background noise ($N(f)$) is statistically independent with the device noise of GEMS ($M(f)$), and glottal excitation signal ($\bar{G}(f)$).

$$E[M(f)N(f)] = 0 \tag{54}$$

$$E[G(f)N(f)] = 0 \tag{55}$$

Therefore, we can derive the statistical independency property between the background

noise and the signal acquired from GEMS

$$E[\bar{G}(f)N(f)] = H(f) \cdot E[G(f)N(f)] + E[M(f)N(f)] \quad (56)$$

$$= 0 \quad (57)$$

From equation (52) and (53), the clean speech signal can be described as:

$$S(f) = \frac{T(f)}{H(f)}\bar{G}(f) - \frac{T(f)}{H(f)}M(f) \quad (58)$$

Denote $\bar{T}(f) = \frac{T(f)}{H(f)}$, equation (51) becomes

$$X(f) = \bar{T}(f)\bar{G}(f) - \hat{T}(f)M(f) + N(f) \quad (59)$$

Multiply $\bar{G}^*(f)$ to both sides of equation 59 and take the statistical expectation, we will have

$$E[X(f)\bar{G}^*(f)] = E[\bar{T}(f)\bar{G}(f)\bar{G}^*(f)] - E[\bar{T}(f)\bar{G}(f)M(f)] + E[N(f)\bar{G}^*(f)] \quad (60)$$

The last two terms on the right side equal to zero. Therefore, $\bar{T}(f)$ can be estimated as:

$$\bar{T}(f) = \frac{P_{gx}(f)}{P_{gg}(f)} \quad (61)$$

As observed in the signal acquired by GEMS, the device noise $M(f)$ is very small. In this case, the clean speech can be estimated as

$$\hat{S}(f) = \bar{T}(f)\bar{G}(f) \quad (62)$$

4.3.2 Filter Implementation

Fig. 14 indicates the analysis of GEMS signal alignment with corresponding acoustic signal.

Therefore, the diagram of GCORR algorithm is proposed as shown in Fig, 15.

4.3.2.1 Synchronization

The GEMS signal is measured at or near the point of vocal tract excitation. Before the same signal is measured by the acoustic microphone, it must travel through the vocal tract, out the mouth, and some distance to the microphone. Thus, the acoustic signal is delayed relative to the excitation signal and this condition may be exacerbated by delays in the

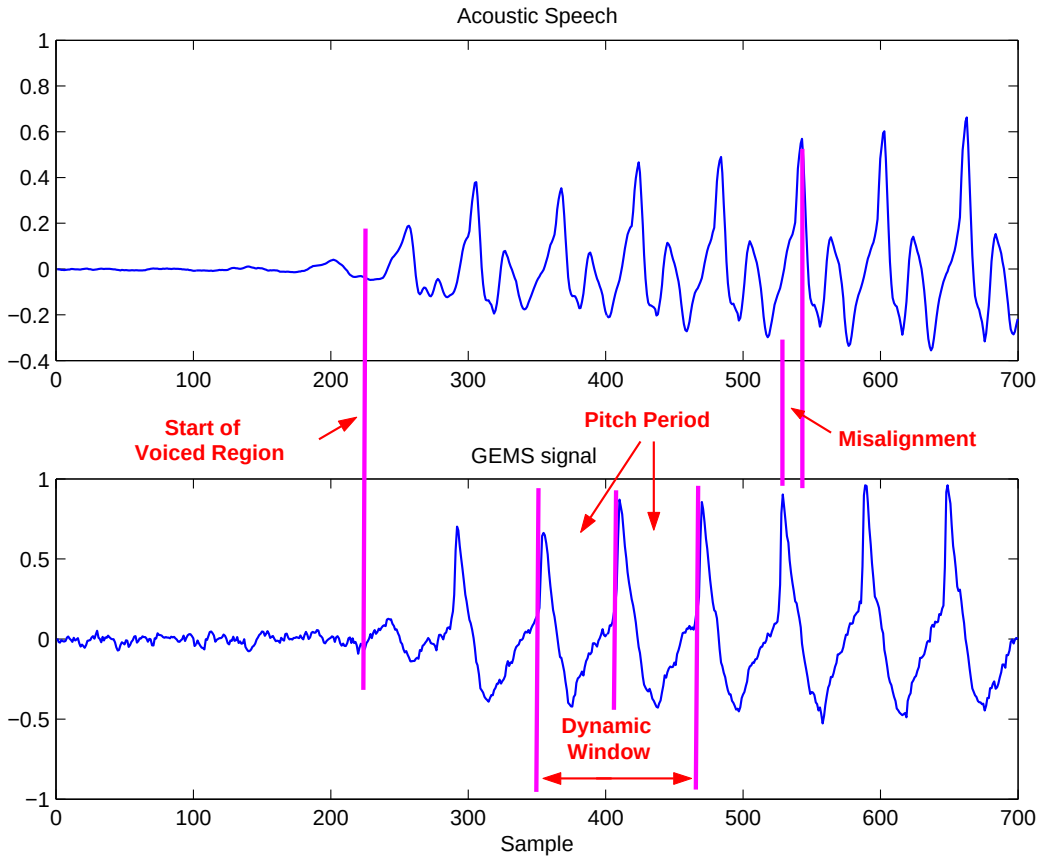


Figure 14: Analysis of GEMS signal alignment with acoustic speech.

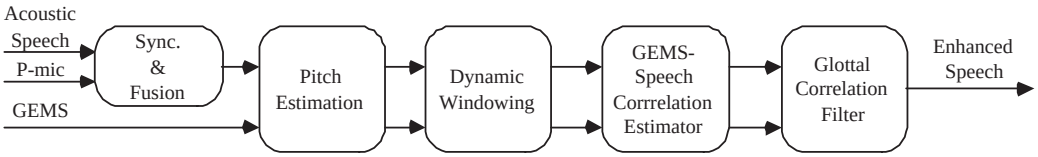


Figure 15: The diagram of GCORR filter.

signal acquisition systems. An example of the resulting misalignment is shown in Figure. 2 but it should be noted that this is only an example and the actual misalignment can vary significantly for different setups. Thus, a synchronization is done to eliminate the delay between them by evaluating cross-correlation of the LPC residual [76] of the speech and the derivative of the GEMS.

4.3.2.2 Pitch Estimation

As demonstrated in Fig. 14, the GEMS signal is quasi-periodic and relatively immune to acoustic disturbance, which makes an accurate pitch estimation easier. A zero-crossing algorithm was introduced by Burnett that is effective in estimating pitch if a very high quality GEMS signal is obtained. However, we have observed that in practice the GEMS signal is sometimes non-ideal, especially when the talker moves. A more robust pitch estimation is applied by carefully finding points of glottal closure using points of maximum negative slope.

4.3.2.3 Dynamic Windowing

Speech signals can only be considered stationary for a short time; however, the Wiener filter is based on an assumption of stationarity. Therefore, in the noise suppression algorithm, the window length should be small enough to make the stationarity assumption valid. However, to achieve good harmonic separation in the spectrum, it is desireable to have as long a window as possible. By selecting a dynamic window that is exactly equal to an integer multiple of the pitch period, it is possible to keep the window short and to improve the harmonic separation sufficiently. A dynamic window with the length of two pitch periods, from the precise start to the end, is shown in Figure. 2. This windowing method helps in increasing the energy compaction of the harmonics in the signal and improves analysis accuracy. The enhanced output is smoothed by overlapping adjacent windows by one pitch period.

4.3.2.4 Filter Construction

As observed, the P-mic signal contains clearer low-pass vocal tract formants than those of normal microphone. The low frequency vocal tract information is important to speech intelligibility, especially for the words with start-consonant of nasal. Therefore, the low-frequency components (<300Hz) of acoustic signals are replaced by corresponding P-mic

Table 8: Pseudocode of GCORR algorithm

Input time-domain signals from acoustic microphone ($x(k)$), GEMS ($g(k)$), and P-mic ($x_p(k)$)
• Synchronize signals by maximizing the cross-correlation
• Estimate the pitch and pitch epoches by zero-crossing
• Construct adaptive frames using the length of two epoches
• For all time frame k
◦ Compute modified speech signal $\hat{X}(f)$ using low-pass fusion with P-mic signal in equation (63)
◦ For all frequency bin f
· Compute GEMS-Speech correlation $P_{g\hat{x}}(f)$ [42]
· Compute the GCORR output $S_G(f)$ using equation (64)

signals to produce more intelligible speech, as illuminated in equation. 63.

$$\hat{X}(f) = L_p(f)\frac{\sum\limits_{f<300} X(f)}{\sum\limits_{f<300} X_p(f)} \cdot X_p(f) + [1 - L_p(f)] \cdot X(f) \tag{63}$$

Where $L_p(f)$ is a lowpass filter with cut-off frequency at 300Hz, $X_p(f)$ and $X(f)$ represent the P–mic and acoustic signals respectively.

The enhanced output of GCORR is constructed by evaluating the statistical properties:

$$S_g(f) = \frac{P_{g\hat{x}}(f)}{P_{gg}(f)} \cdot G(f) \tag{64}$$

Where, $(P_{g\hat{X}}(f))$ represents the cross-correlation of the GEMS signals and the modified acoustic signals [42].

The detail implementation is described in [42]. This approach can also be applied to the input of other sensors, where the acquired signal is highly correlated to clean speech and immune to background noise, such as bone-conduction mic.

4.3.3 Evaluation

The sample spectrograms for GCORR filter are provided in Figure 16 and Figure 17, where the speech data is from the ARCON database, where GEMS, P-mic and other non–air conductive sensors are available.

The figures reveal that the GCORR algorithm produces fine a spectral shape in the enhanced speech even in ultra low SNR cases, especially in low–frequency bands, as seen

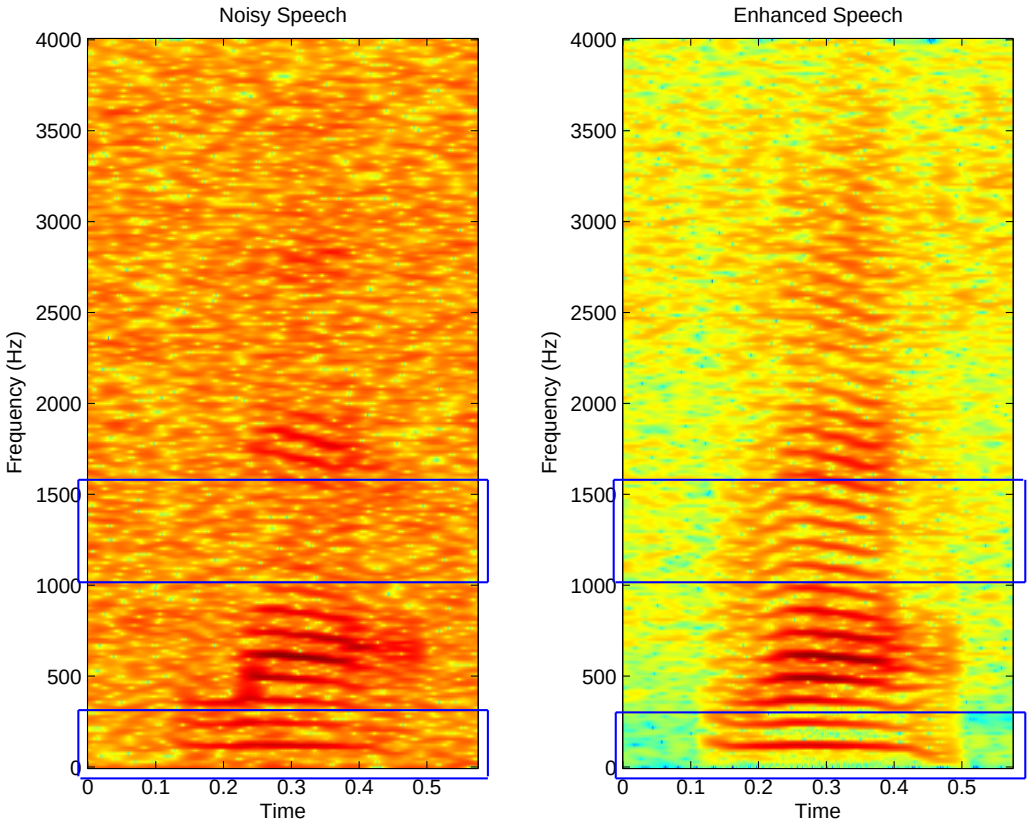


Figure 16: Evaluation of GCORR by speech spectrogram in white noise (SNR=0dB).

from the marked region in the figures. The gain comes from the fact that the non-acoustic sensor yields a better measure in those bands.

4.4 Multi–Sensor Speech Enhancement System

The CQ-GCORR system is a hybrid algorithm [44] that combines the strengths of the biologically inspired (CQ) algorithm and the enhanced Glottal Correlation (GCORR) algorithm. The diagram is shown in Figure 30. The signal fusion criterion is to maximize the energy function weighted from subjective *a*–priori knowledge.

4.4.1 Performance Analysis of Algorithms

From informal subjective measures, the CQ algorithm appears particularly well suited for finding and extracting unvoiced portions of speech from the acoustic signal. This algorithm seems more effective at higher frequencies. GCORR is useful for removing noise and estimating the vocal-tract function; however, in the GEMS signals we have observed, high

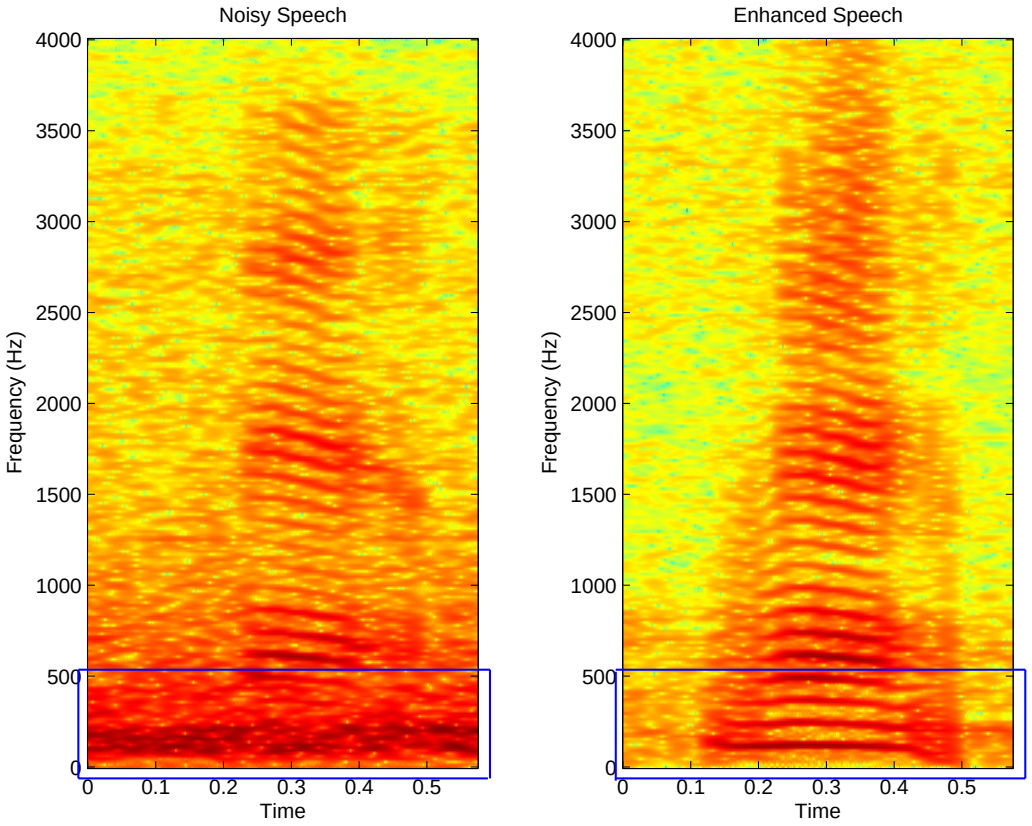


Figure 17: Evaluation of GCORR by speech spectrogram in tank noise (SNR=−5dB).

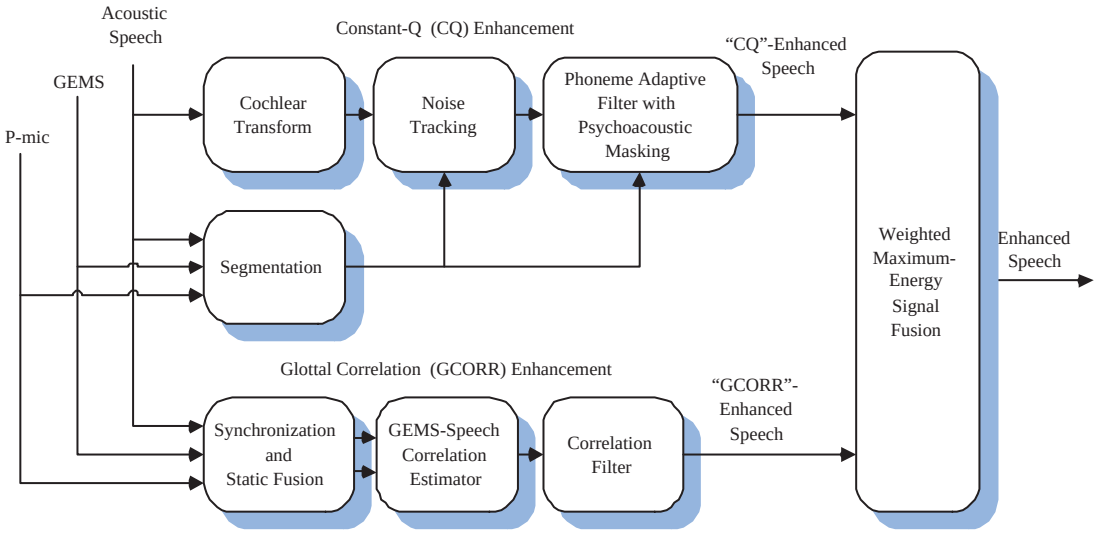


Figure 18: Block diagram of the proposed CQ–GCORR speech enhancement system

frequency information is occasionally lost. Additionally, the mid-frequencies are very sensitive to pitch variation over time and may be a source of artifacts when GEMS signal does

not match the actual pitch perfectly. This is even true for the signals above 500Hz from the experimental analysis. For these reasons it seems more reasonable to use the GCORR algorithm to estimate low frequency portions of speech signals and the CQ algorithm for higher frequencies. Figure 19 indicates the comparison of two algorithms, in term of segmental normalized SNR improvements in each critical sub-band. The segmental normalized SNR improvements were calculated as

$$SNR_{imp} = \frac{1}{L}\sum_{m=0}^{L-1} 10 \cdot \log_{10} \frac{\frac{1}{N}\sum_{k=0}^{N-1} n^2(m,k)}{\frac{1}{N}\sum_{k=0}^{N-1}\left[s(m,k) - \hat{s}(m,k)\right]^2} \tag{65}$$

where L represents the number of frames in the voiced signal and N is the number of samples in mth frame. All the signals, $n(m,k)$, $s(m,k)$, and $\hat{s}(m,k)$, are normalized by the same factor, in order to get the equalized evaluation. The comparison was expressed by the SNR improvements gains of GCORR over CQ.

$$DSNR_{imp} = SNR_{imp}(GCORR) - SNR_{imp}(CQ) \tag{66}$$

4.4.2 Maximum Energy (ME) Signal Fusion

From the results, GCORR performs much better in low frequency bands (critical band ¡ 5), and low input SNR conditions. But CQ outputs are a little better in other cases. Therefore, at the back-end, a signal fusion is performed to merge strengths of both outputs to improve speech intelligibility.

In the signal fusion, the observed signals include the output of CQ ($\hat{s}_c(k)$), the output of GCORR ($\hat{s}_g(k)$), and the noisy signal ($x(k)$). We propose the final output of CQ-GCORR system is constructed as the linear combination of the observed signals in all sub-bands.

$$\hat{s}(k) = \sum_{m=1}^{M}\left[\hat{h}_c(m)\hat{s}_c(m,k) + (1-\hat{h}_c(m))\hat{s}_g(m,k)\right] \tag{67}$$

where $\hat{s}_c(m,k)$ and $\hat{s}_g(m,k)$ represent the sub-band signal of CQ and GCORR output respectively.

The objective of is set to minimize the mean square error (MSE), which can be described

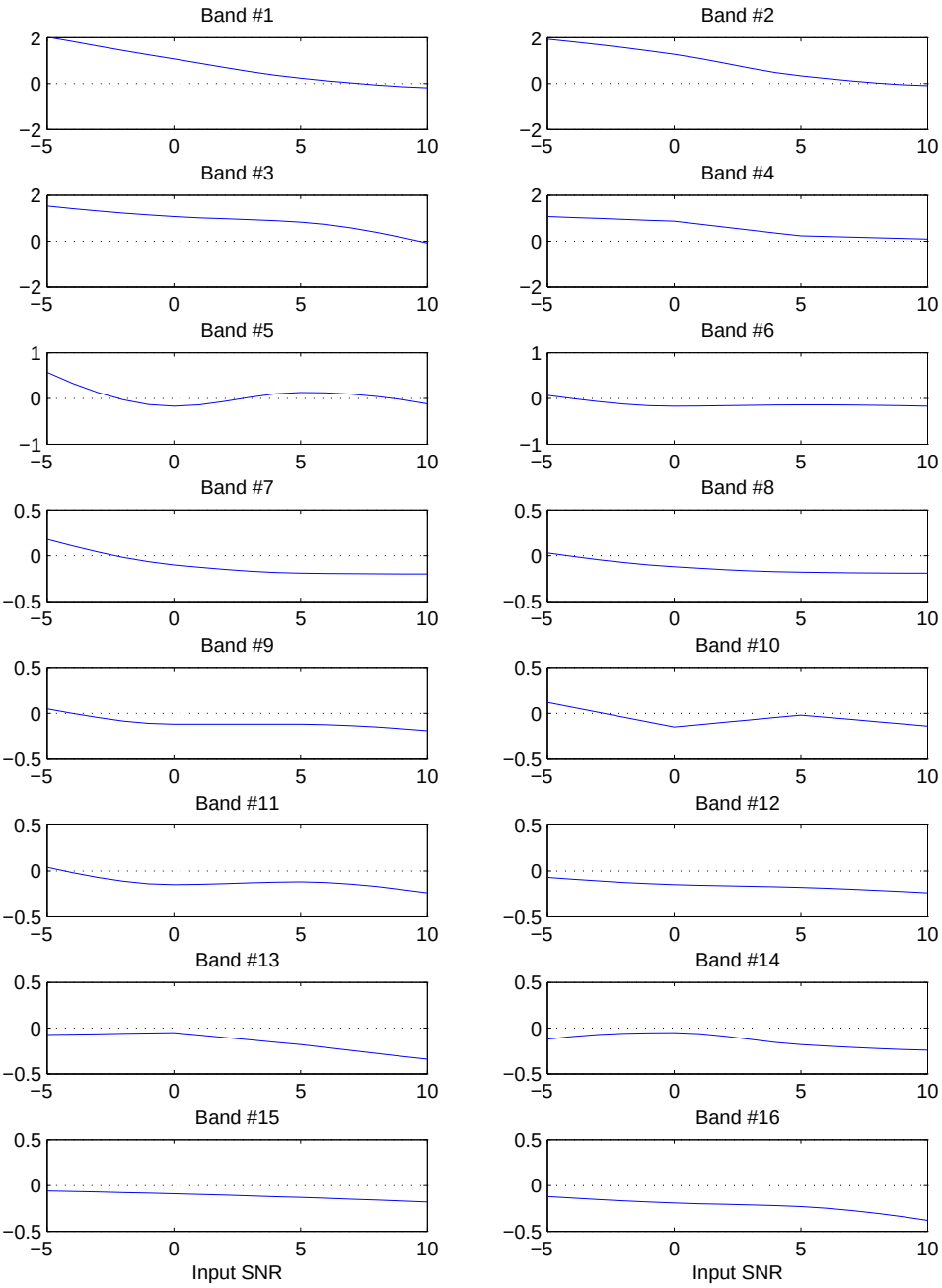


Figure 19: The comparison of normalized SNR improvements in critical subband between GCORR and CQ

as:

$$min \sum_{k} \left(s(m,k) - \hat{s}(m,k)\right)^2 = \arg\min_{h_c(m)} \sum_{k} \left(s(m,k) - \hat{h}_c(m)\hat{s}_c(m,k) - (1 - \hat{h}_c(m))\hat{s}_g(m,k)\right)^2 \tag{68}$$

There is no explicit solution for this equation. Some *a prior* knowledge is needed to approach the optimal fusion. In this implementation, we look for a solution with lower computation complexity in order to process the sinal in real-time applications. The minimization of equation 68 equals to the maximization of equation 65. Therefore, if we set the observation as the ratio of the smoothed signal envelope of CQ output ($\hat{S}_c(m,k)$) and GCORR output ($\hat{S}_g(m,k)$).

$$\lambda(m,k) = 10 \cdot \log_{10} \frac{\hat{S}_c(m,k)}{\hat{S}_g(m,k)} \tag{69}$$

The maximization of equation 65 can be approximated as:

$$\arg\max_{h_c(m)} (SNR_{imp} | \lambda(m,k)) \tag{70}$$

where the parameters $h_c(m)$ equals to the conditional probability of CQ outperforming GCORR given the observation $\lambda(m,k)$.

$$h_c(m) = p(DSNR_{imp} < 0 | \lambda(m,k)) \tag{71}$$

The conditional probabilities were determined in off-line training by the clean speech with additive noise. In the implementation, interpolation and smoothing were used to avoid the transitional variations.

4.4.3 Illustration of CQ-GCORR behavior

Figure 20 shows the behavior of the CQ-GCORR system in M2 fighting vehicle noise environment. The GEMS signal exhibits large attenuation of background noise. But it can not capture the speech in unvoiced section (phoneme "t"). This factor gives the same problem in the output of GCORR algorithm. The combination of CQ and GCORR indicates better performance.

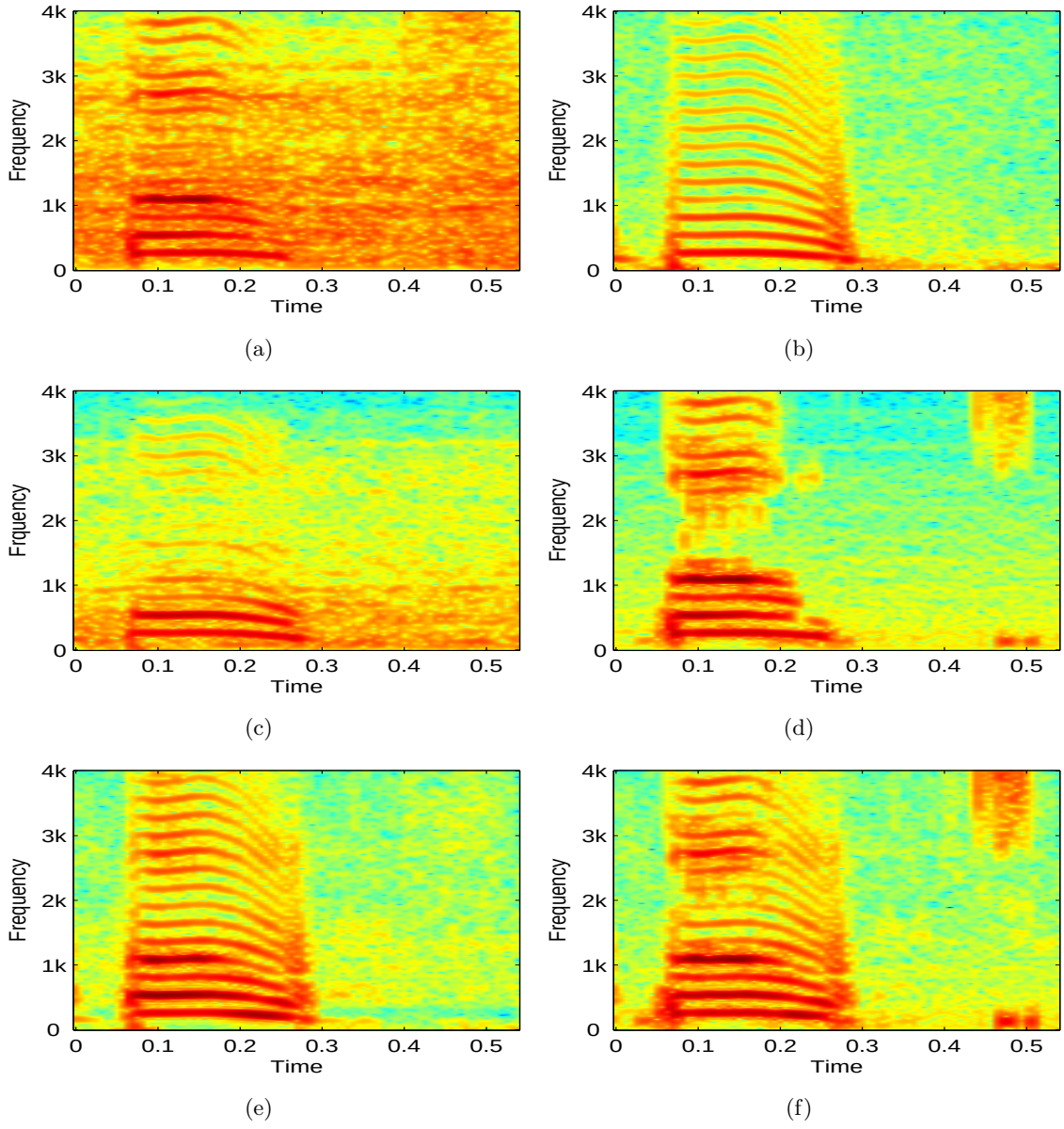


Figure 20: Results of a speech clip (word "boot") in M2 fighting vehicle noise environment ($SNR = 5dB$), (a) noisy speech, (b) the GEMS signal, (c) the P-mic signal, (d) the CQ output, (e) the GCORR output, (f) the CQ–GCORR output.

4.5 Performance Evaluation

This section presents the performance evaluation of the proposed speech enhancement system. Speech data were selected from arcon corpus, an extensive multi-sensor speech corpus collected from ten male and ten female talkers in nine different acoustic noise environments. The sensors consisted of the introduced GEMS, P-mic. In the tests, three types of noise in

harsh environment were used: M2 Bradley Fighting Vehicle, MOUT (military operations in urban terrain), and Blackhawk helicopter. For the experiments, the signals were down–sampled to 8kHz. Noise has been added to the clean speech signal with a varying SNR. We choose the well–known EMSR algorithm as the base–line system to make comparisons.

The performance evaluation is based on the application of objective or subjective quality and intelligibility measures. Generally, two main undesirable effects are produced and concerned by research in speech enhancement: residual noise and speech distortion. To particularly observe the structure of residual noise, it is important to analyze the time-frequency distribution of the enhanced speech. In this work, speech spectrograms are presented. Subjective tests for the low-bit-rate-coded speech (2400bps) were performed to verify the improvement in speech intelligibility and quality. Finally, in order to further validate the subjective evaluation, objective measures, including segmental SNR improvement and log spectral distortion, were provided.

4.5.1 Intelligibility Assessment

The diagnostic rhyme test (DRT) [34, 59] uses monosyllabic English words that are constructed from a consonant-vowel-consonant sound sequence. In the DRT, one hundred and ninety two words are arranged in ninety-six rhyming pairs which differ only in their initial consonants. Listeners are shown a word pair, then asked to identify which word is presented by the talker. Carrier Sentences are not used. The DRT is based on a number of distinctive features of speech, and its test results reveal errors in discrimination of initial consonant sounds.

The DRT is a quite widely used method. It provides lots of valuable diagnostic information how properly the initial consonant is recognized. The score is highly correlated to speech intelligibility and accepted as standard indication. The results are provided in Table 9 and 10. The difference in DRT scores are:

- Average increase of 3.46 points for all speakers
- Average increase of 4.45 points for females

- Average increase of 2.69 points for males
- Biggest improvement for voicing (+13.07)
- Slight improvements for Nasality, sustenation, EXP
- Slightly worse for graveness, sibilation, compactness

Although some small degradation in three feature sets, in average, the CQ-GCORR system gives significant improvement in speech intelligibility for low-bit-rate speech coding in acoustic harsh environments.

Table 9: DRT scores of the enhanced speech signals in harsh environments (SNR<5dB) in low-bit-rate (MELP@2400bps) coding.

Noise Type	Speaker	EMSR	CQ-GCORR	Gain
Blackhawk Helicopter	Male	77.82	78.52	**0.70**
	Female	80.62	85.51	**4.89**
M2 Bradley vehicle	Male	67.36	70.53	**3.17**
	Female	76.43	82.42	**5.99**
Gun Shoot	Male	80.99	85.20	**4.21**
	Female	87.67	90.02	**2.35**

Table 10: DRT scores of different feature sets

Feature Set	EMSR	CQ-GCORR	Gain
Voicing	74.45	87.52	**13.07**
Nasality	80.62	86.41	**5.79**
Sustenation	64.36	66.53	**2.17**
Sibilation	79.83	78.92	**-0.91**
Graveness	75.52	75.20	**-0.32**
Compactness	88.43	85.12	**-3.31**
EXP	77.07	80.02	**2.95**

4.5.2 Quality Measure

Pair comparison method is usually used to test system overall acceptance. The tests were performed with the output of a coder using CQ–GCORR as a front-end noise suppressor over standard CELP. The processed signals were collected in low UH-60A Blackhawk helicopter

Table 11: Pair comparison A/B test results, Percent Preference for CQ–GCORR over CELP, in low UH-60A Blackhawk Helicopter noise environment

Talker	Score	Gender	Overall
Female–1	85.42%		
Female–2	68.75%		
Female–3	66.67%	73.61%	
Male–1	60.42%		
Male–2	41.67%		
Male–3	64.58%	55.56%	64.58%

noise cases (SNR$>$ 10dB). Continuous Harvard BP sentences were used instead of the isolated words in DRT. The percentage of preference of CQ–GCORR in Table 11 indicates that the proposed system produces enhanced speech of higher quality, in other words, it suppresses more noise and fewer residual artifacts remain.

The amount of noise suppression is also demonstrated by the segmental SNR improvement in the voiced segments. The results are shown in Table 12. The proposed system increases the noise suppression levels in all noise conditions.

Table 12: Segmental SNR improvement for voiced segments in various noise conditions

Noise Type	Input Seg. SNR (dB)	Seg. SNR Improvement (dB)			
		EMSR	CQ	GCorr	C-G
Blackhawk Helicopter	-5	8.04	9.09	9.35	**11.67**
	0	7.61	8.42	8.64	**10.72**
	5	7.08	7.91	7.65	**9.47**
M2 Bradley Vehicle	-5	5.94	8.10	9.01	**11.17**
	0	5.57	7.69	8.45	**10.27**
	5	5.24	7.11	7.42	**9.13**
Gun Shoot	-5	3.17	5.15	6.89	**7.27**
	0	2.81	4.75	6.60	**6.89**
	5	2.24	4.09	6.16	**6.27**

The log spectral distance (LSD) is computed in the Fourier domain. This is used instead of the computation of SNR in time domain, which will not produce meaningful results as the phase of the estimated high frequency components will not match that of the original 8kHz signal. The average log spectral distance, LSD, between the original speech, S, and

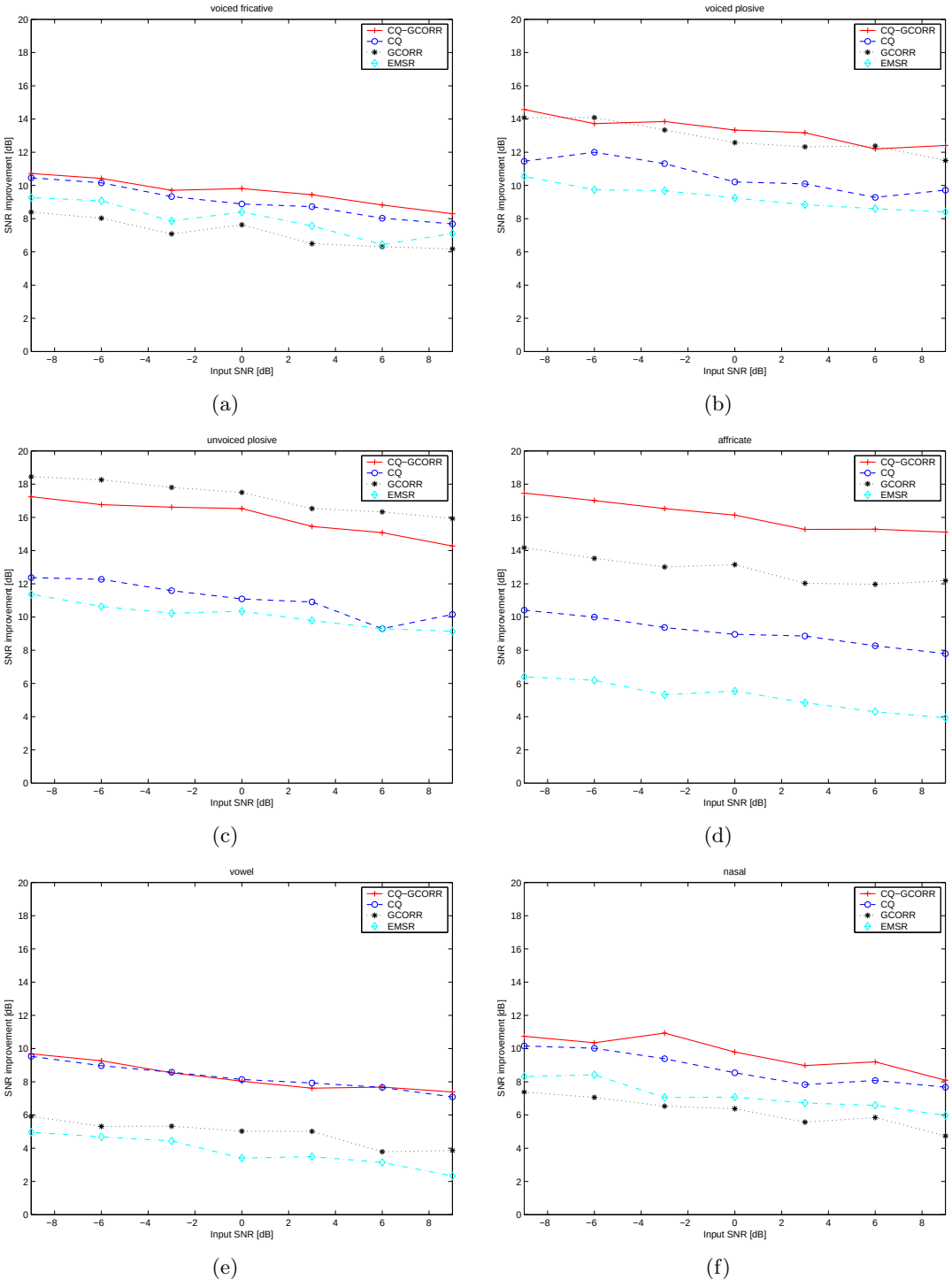


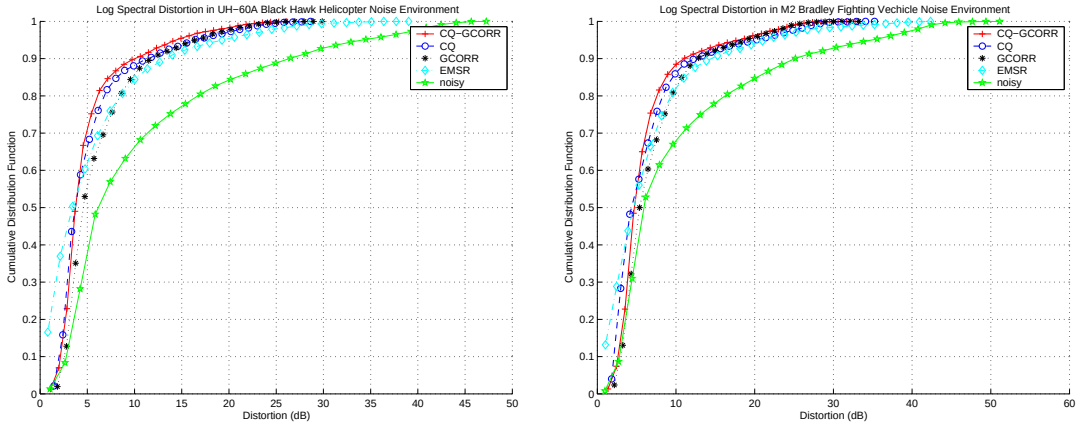
Figure 21: SNR Improvement for each phoneme class in M2 Bradley fighting vehicle noise environment. (a) voiced fricative; (b) voiced plosive; (c) unvoiced plosive; (d) affricate; (e) vowel; (f) nasal. The tested methods are the following: (+) CQ–GCORR, (o) CQ, ($*$) GCORR, ($\diamondsuit$) EMSR.

the enhanced speech, X, is defined as:

$$LSD = \frac{1}{K} \sum_{k}^{k=K-1} \left[\frac{1}{\omega_s} \int_{-0.5\omega_s}^{0.5\omega_s} \left(\log_e |X(\omega)| - \log_e |S(\omega)| \right)^2 d\omega \right]^{0.5} \tag{72}$$

where K is the number of frames.

The distortion measures are plotted as cumulative distribution functions of the magnitude of speech distortion. From overall results in Fig.22 and analysis of performance on different phonemes in Fig.23, the enhanced output of CQ–GCORR system suffers less speech distortion, which can be indicated from higher cumulative distribution up to large distortion magnitude in the curves. As observed, for some phonemes, like unvoiced plosive, the log spectral distortions of CQ and GCORR are higher than that of EMSR. With the advantage of prior–knowledge–based maximized energy signal fusion, CQ–GCORR takes the "significant" signals of each outputs in different frequency bands. The spectral distortion is still superior to EMSR.



(a) (b)

Figure 22: Overall log spectral distortion of enhanced outputs in high noise environment (SNR=0dB). (a) Blackhawk helicopter noise; (b) M2 Bradley noise. The measured signals are the following: the enhanced output of (+) CQ–GCORR, (o) CQ , ($*$) GCORR, ($\diamondsuit$) EMSR, and (x) noisy speech.

4.6 Summary

General speech enhancement algorithms utilize the inputs from acoustic sensors only. Although many optimization techniques have been developed, the enhanced outputs may still suffer from considerable speech distortion, especially in low SNR conditions. As a result, the

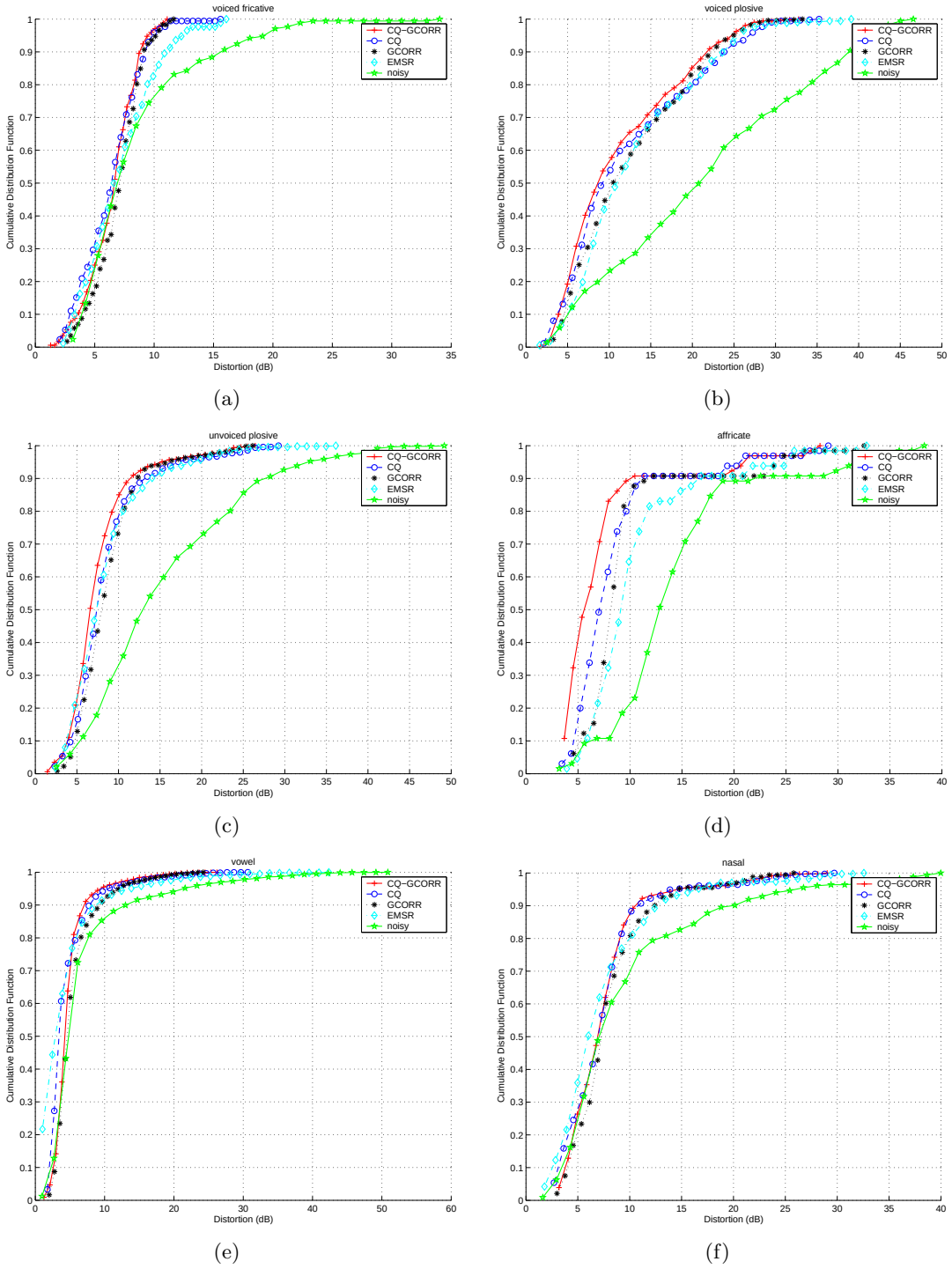


Figure 23: Log spectral distortion measure enhanced outputs in M2 Bradley fighting vehicle noise environment (SNR=0dB). (a) voiced fricative; (b) voiced plosive; (c) unvoiced plosive; (d) affricate; (e) vowel; (f) nasal. The measured signals are the following: the enhanced output of (+) CQ–GCORR, (o) CQ , ($*$) GCORR, ($\diamondsuit$) EMSR, and (x) noisy speech.

speech intelligibility is degraded. In the proposed system, we exploit the state-of-the-arts non-acoustic sensors and introduce a hybrid speech enhancement system using perceptually inspired techniques. Both subjective and objective experiments were conducted and compared to the EMSR algorithm in various noise types and levels. The proposed system is effective in suppressing background noise. Significant improvement of speech intelligibility for low-bit-rate speech coding in harsh environments is achieved.

CHAPTER 5

SPEECH BANDWIDTH EXTENSION

In many speech transmission systems, such as the digital public telephone system, low-bit-rate-speech coding environment (MELP), the bandwidth of speech is limited to 4KHz. This kind of speech is so called "telephone speech". Compared to natural speech, telephone speech has a significantly degraded performance. The bandwidth limitation of telephone speech reduces speech intelligibility by about 10 percent, and decreases the subjective quality score, which is measure in terms of the subjective mean opinion score (MOS) by more than one point [56].

Owing to the importance of the acoustic bandwidth for speech intelligibility and especially for subjective quality. It is worthwhile to extend the speech bandwidth. Particularly, in digital communication and hands-free telephony, there is a demand for enhancing the subjective speech quality.

To avoid the modification of narrowband communication systems, where the receiver does not have the access to the wideband signal, recent work [26, 38, 19, 46, 88] has been done to artificially extend the narrowband speech to wideband speech by signal processing techniques. These techniques are motivated from the fact that the spectral envelope of the lower and upper frequency bands of the speech signal are dependent, which can be illuminated from the speech production model. In most real-world conditions, the bandwidth of speech signals is corrupted by background noise or transmission media, such as telephone line, VoIP, and low-bit-rate speech coders. Bandwidth extension methods are used to recover the lost information based on the redundancy nature of human speech. The framework, as shown in Figure 24, consists of two major components: residual extension and envelope extension.

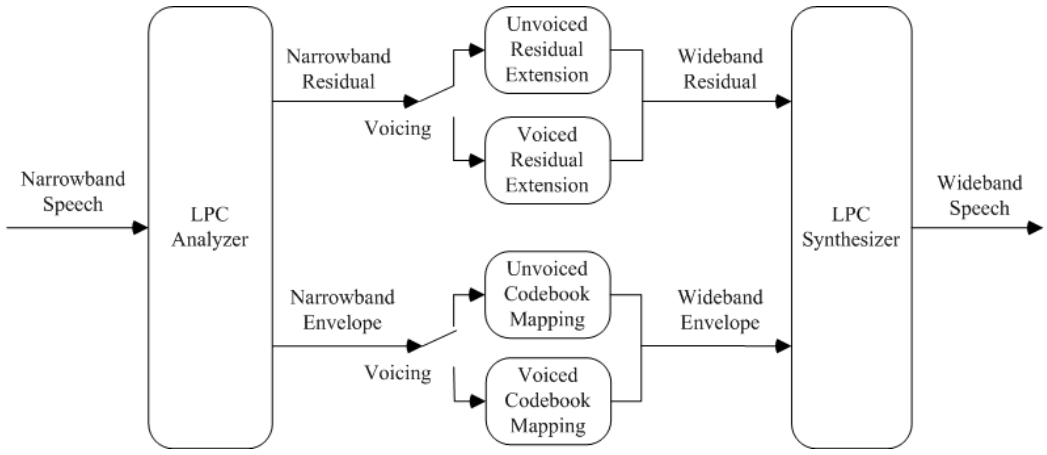


Figure 24: The basic framework of bandwidth extension.

5.1 Redundancy Between Frequency Bands

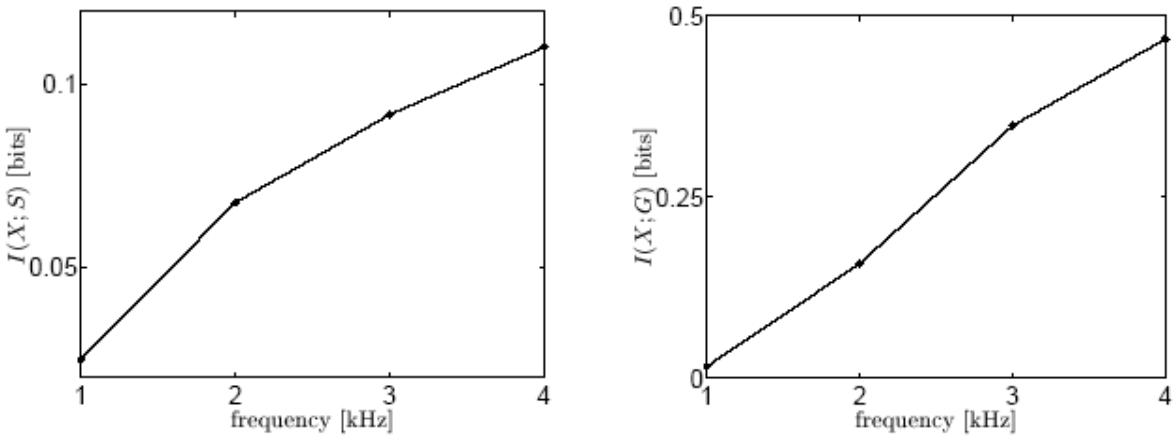
The motivation for the bandwidth extension methods is the fact that the spectral envelope of the lower and higher frequency bands of the speech signal are dependent, i.e., the low band part of the speech spectrum provides information about the spectral shape of the high band part. This results from speech being created by a physical source.

If the logarithmic spectral energies of frequency bands have a Gaussian distribution, their relation can be described with a correlation function. However, it is well known that the logarithmic spectral energies of the speech signal are non-Gaussian and thus have statistical moments of order higher than two, which would not be accounted for with a correlation measure. In this case, mutual information is an appropriate measure.

The redundancy analysis attempts to investigate the amount of information that is shared between the low and high band in speech. The objective is to determine an approximate value of the mutual information between the high band and various widths of the low band of spectral envelopes of speech. We have in this paper only considered the mutual information between spectral envelopes. The results should provide information on whether there is a frequency region in the low band, that contains almost all information about the high band. If there is, we could claim that speech coders coding the high band independently of the low band are wasting bits in representing something which is predictable from the low frequency band. In the analysis work, only the slope respectively the gain of the

high band spectral envelope are used to capture the behavior of the high band. The slope of the high band conveys partial information on whether a speech segment is voiced (v) or unvoiced (uv). A v/uv decision can also be made from a low frequency band 0 - 4 kHz. This suggest that the low band contains information about the high band slope in the same order of magnitude as the entropy of a v/uv classification. If we assume 80% of the speech to be voiced, this results in an entropy of 0.72 bit. If the slope contained full information about the v/uv classification, 0.72 bit would be a lower bound on the mutual information between the slope and the low band. However, we do not suggest that this is the case since only partial information on a voiced/unvoiced classification is conveyed by the slope. Looking at the LPC spectrum at the high frequency band we can see that there are some peaks and valleys, which will not be captured using only the slope and gain. However, this work is a first step in the direction of finding the true mutual information between the low and high frequency bands.

A lower bound estimation on the mutual information between frequency bands was introduced in [73].



(a) (b)

Figure 25: (a) A lower bound on the mutual information between the slope of the high band given spectral envelope representation of the low band for regions 0-1, 0-2, 0-3 and 0-4 kHz. (b) A lower bound on the mutual information between the gain of the high band given spectral envelope representation of the low band for regions 0-1, 0-2, 0-3 and 0-4 kHz.

The data set consisted of 2200 speech files (sampled at 16 kHz) from the TIMIT data base, yielding 600000 segments of length 20 ms using 50 % overlap. The Log-Area-Ratio

(LAR) was used to represent the slope parameter:

$$s = \log_{10}(\frac{1-l}{1+l}) \tag{73}$$

where l is the first reflection coefficient. The reflection coefficient was calculated from a first order LPC analysis. From the same analysis the amplification b of the LPC filter $A(z)$ was determined assuming a unit variance input:

$$A(z) = \frac{b}{1 - lz^{-1}} \tag{74}$$

The logarithm of b is used as the gain parameter:

$$g = log_{10}(b) \tag{75}$$

From each speech file, high band and low band speech files were created. The low band speech file contained frequencies up to P kHz, where P ranged from 1 to 4 kHz. The high band speech file was created by first high-pass filter- ing the speech signal at a cut-off frequency of 4 kHz. The high-pass filtered signal was then modulated with a cosine to move the signal to the band 0 - 4 kHz, low-pass filtered at 4 kHz and finally down-sampled by a factor 2. For each speech segment the slope and the approximate MMSE estimate of the slope were found and the estimation noise was determined.

Observing Figure 25(a), we see that there is no less than 0.1 bit of mutual information between the spectral envelope of the low band frequency region 0 - 4 kHz and the slope of the high band. From Figure 25(a) we see that largest increase in mutual information is achieved when we increase the information about the low band from representing 0 - 1 kHz to representing 0 - 2 kHz. The mutual information then seems to level out as more information about the spectral characteristics of the low band are given. One possible explanation is that we have one formant in the region 0 - 1 kHz from which alone it is hard to estimate the slope, since the total number of formants determine the slope of the spectrum at the high band, assuming an all-pole signal model. All-pole models form a good model of the vocal tract, and, thus, of the spectral envelope. However, by extending the region to 0 - 2 kHz we have a significantly better estimate of how many formants there are in the region 0 - 4 kHz and thus we obtain a better prediction of the slope.

Figure 25(b) shows the results obtained when we used the gain G instead of the slope S in the simulations. From Figure 25(b) we can see that the there is no less than 0.45 bit of mutual information between the gain and the spectral envelope of low band frequency region 0-4 kHz. The curve indicates that the shared information between the low band and the gain of the high band is related to how much of the speech signal energy we observe.

5.2 BWE Framework

Within the state-of-the-art speech bandwidth extension techniques, codebook mapping is a commonly used and promising method. In this thesis,we present our speech bandwidth extension system using improved codebook mapping towards increased phonetic classification based on *a prior* knowledge.

The proposed speech bandwidth extension system [45], shown in Figure 26, consists of two major modules: excitation extension and spectral envelope extension. The received narrowband speech signal is first analyzed, the narrowband spectral envelope representatives, line spectral frequencies (LSFs) are obtained. The residual of the analyzer is the narrowband excitation. Then, the excitation and spectral envelope of the narrowband are extended to wideband using corresponding extension model. In the synthesis part, The extended wideband excitation and spectral envelope are synthesized. The output is passed through a high-pass filter, and is summed with the narrowband speech, which is upsampled by a factor of 2 and passed through a low-pass filter.

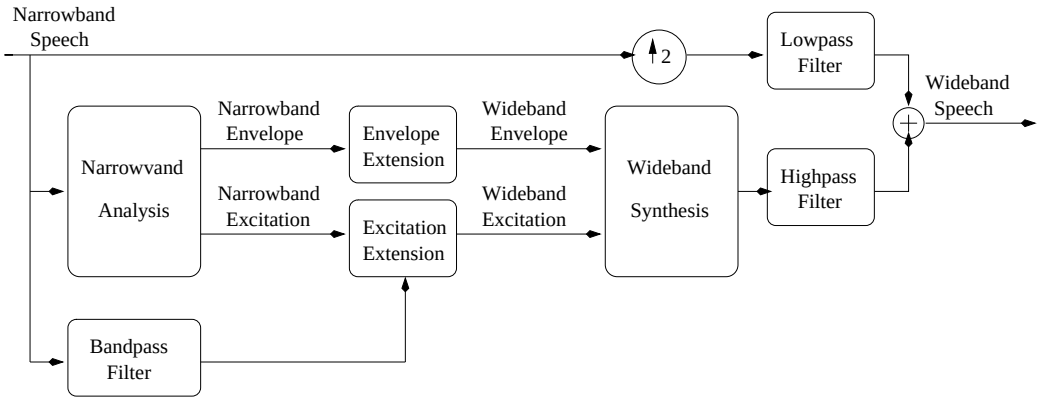


Figure 26: Block diagram of the proposed system

5.2.1 Excitation Extension

In wideband coding technology, pitch adaptive modulation has been shown to provide better wideband excitation. This was confirmed in speech bandwidth extension in [78]. The performance is better than spectral folding and spectral replication. In our excitation extension module, we use the envelope of the bandpass signal between 2.5-3.4KHz to modulate the wide-band excitation.

the Linear Prediction (LP) residual of voiced phonemes contains weak pitch harmonics and noise-like components over 4 kHz, while the residual below 3.5 kHz shows strong pitch harmonics.

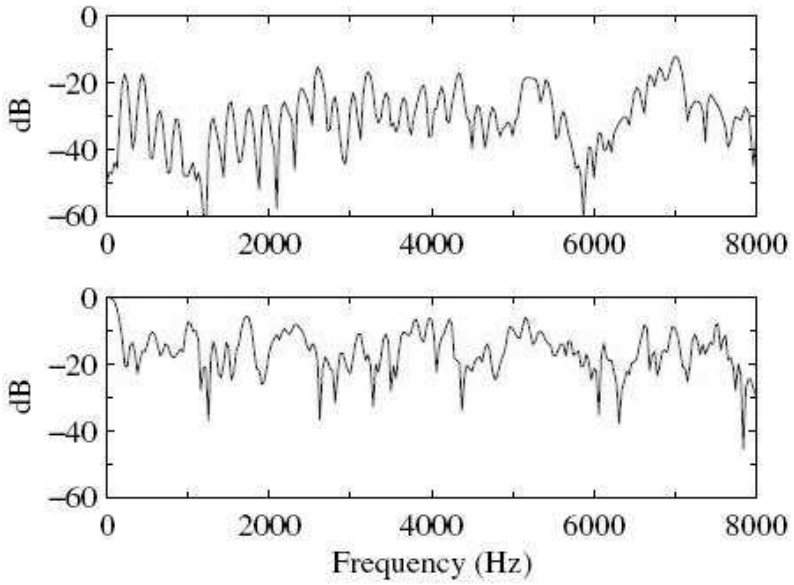


Figure 27: The LP residual spectrum of a voiced phoneme (upper trace) and the LP residual spectrum of an unvoiced phoneme (lower trace).

The block diagram for generating BP-MGN is shown The upsampled narrowband speech passes through a 2C3 kHz bandpass filter. The bandpass signal is

$$s_{bp}(n) = s_{bb}(n)cos(2\pi f_o n). \tag{76}$$

where $f_o = 2.5$ kHz and $s_{bb}(n)$ is a baseband signal. The envelope of the bandpass signal is $\|s_{bp}(n)\|$. The spectrum of the envelope is $S_{bpe}(\omega)$,

$$\mathrm{S_{bpe}}(\omega) = \mathrm{S_{bp}}(\omega)\mathrm{S_{bp}}(\omega) \tag{77}$$

The BP-MGN excitation, $e(n)$ is a bandpass-envelope modulated by a Gaussian noise. The spectrum of the BPMGN excitation $E(\omega)$ is the convolution of the Gaussian noise spectrum $G_n(\omega)$ and S_{bpe} in frequency domain,

$$E(\omega) = G_n(\omega) \cdot S_{bpe}(\omega) \tag{78}$$

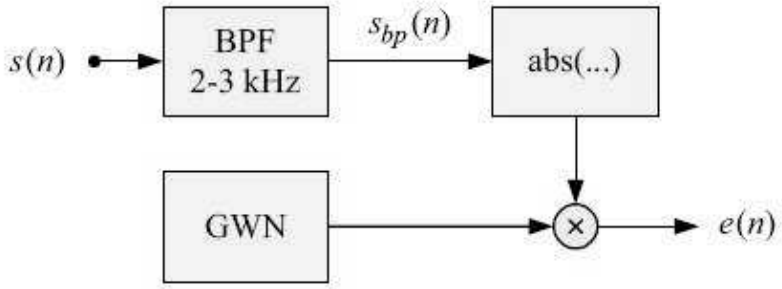


Figure 28: BP-MGN excitation generation block

Figure 29 (top) shows the spectrum of the bandpass signal, $S_{bp}(\omega)$ of a voiced phoneme. It has strong pitch harmonics. Figure 29 (middle) gives the spectrum of the bandpass-envelope, showing the presence of pitch harmonics. Figure 29 (bottom) shows the BP-MGN spectrum, which will be used in the high frequency region.

5.2.2 Spectral Envelope Extension

A number of techniques for estimating wideband spectral envelope have been recently proposed [26, 38, 19, 46, 88]. In particular, codebook mapping is popular used and promising. The principle of this class of algorithms is based on the observation that there occur only a limited number of the typical sounds in speech signals. Accordingly, the codebook mapping approach is based on a pair of coupled codebooks that contain representations of the spectral envelopes of the narrowband and wideband speech, respectively. For each signal frame, the spectral envelope of the narrowband speech signal, represented by the feature vector x is compared to a list of typical narrowband spectral envelopes that are stored in a pre-trained primary codebook. The closet codebook entry is selected. In parallel to the searched primary codebook, there exists a second codebook, which contains corresponding

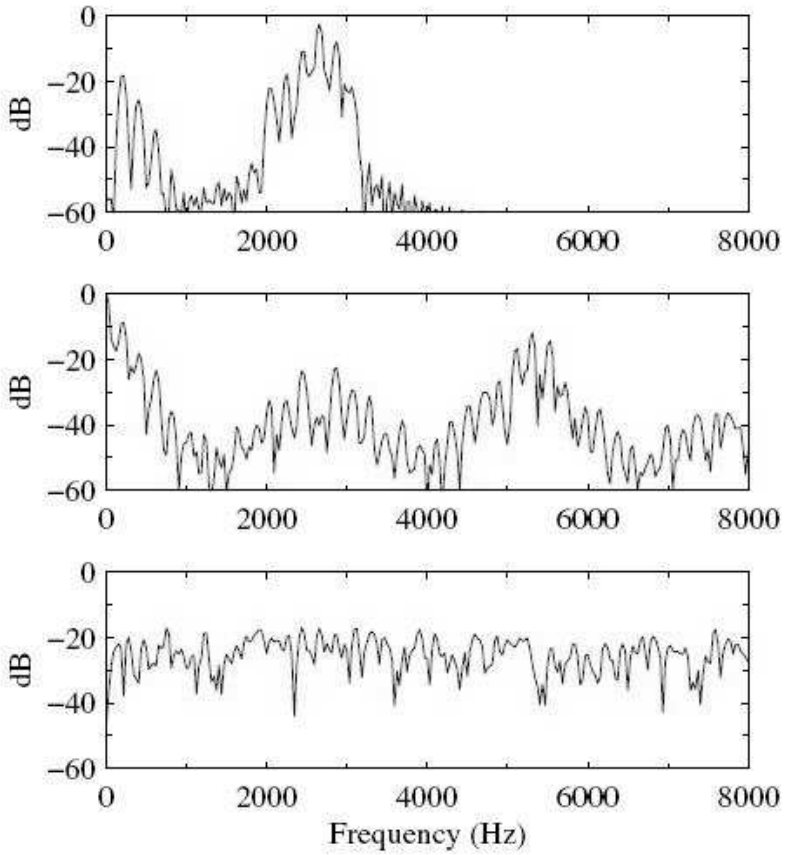


Figure 29: The spectrum of the bandpass signal (top); the spectrum of the bandpass-envelope (middle); BP-MGN spectrum.

wideband spectral envelope representatives. The estimate $\hat{y}$ of the wideband spectral envelope is the entry of the second codebook that is assigned to the selected entry of the primary codebook.

There are many representatives for spectral envelope, including linear prediction coefficients (LPC), Mel-frequency cepstral coefficients (MFCC), LSF and so on. In this application, the spectral envelope is stored as LSF. LSF coefficients describe the spectral envelope in a more direct way. The range of $(0, \pi)$ corresponds proportionally to the whole frequency range of the spectrum. Additionally, LSF has some desirable properties: when LSF values fall in the range $(0, \pi)$, the recovered LPC filter has guaranteed stability; local errors of LSF values only cause local spectral distortion. Therefore, codebook mapping based on LSF values is more tolerant to estimation errors, as a single error cannot harm the whole spectral envelope. A comparison between LPC and LSF has been made in [88]. More important,

the selection of LSF makes it easier to implement an improved codebook mapping method, distance measure based on *a prior* knowledge. This will be covered in later section.

5.2.3 Assessment

It is well-known that the synthesized speech by bandwidth extension provides higher intelligibility, especially for the fricatives where the most spectral energy locates in high-band, and better quality also.

The performance of the speech bandwidth extension using codebook mapping depends on many factors, in which codebook size and codebook partition method (distance measure) are crucial. Log spectral distortion (LSD) is an useful for assessing the performance. The LSD in overall bands (narrowband and high-band) indicates the improvement of a bandwidth extension relative to narrowband speech. To provide a comparison between different bandwidth extension methods, the high-band LSD (HB-LSD) is used, since the actual narrowband signal is given.

Recent research has found that the codebook mapping performance in the speech bandwidth extension saturates for codebook sizes greater than about 256 [46]. Therefore, 256-level codebooks are used in our system. Both narrowband and high-band parameters were 12th-order LSF in the implementation of conventional codebook mapping.

In the preliminary experiments, a speech bandwidth extension system using conventional codebook mapping is implemented. The overall LSD between the synthesized wideband speech and original narrowband speech is indicated in Table 18.

Table 13: The performance of speech bandwidth extension using conventional codebook mapping

	Narrowband	BWE
Overall LSD (dB)	11.45	**5.32**

5.3 Improved Codebook Mapping

In the proposed system, we apply three techniques to improve the performance of codebook mapping: a distance measure towards increased phonetic classification, marginal LSF

interpolation, codebook mapping with memory, and codebook interpolation.

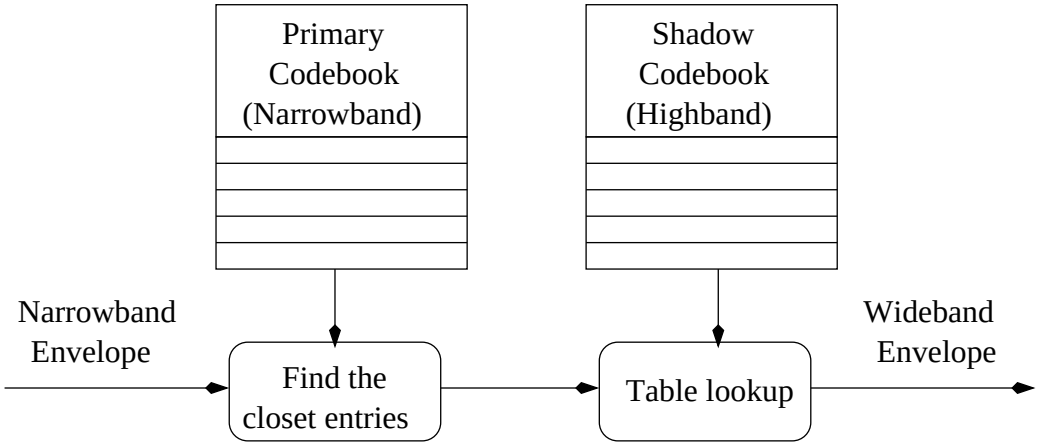


Figure 30: Block diagram of the conventional codebook mapping method

5.3.1 Codebook Training Towards Increased Phonetic Classification

The objective for codebook mapping in bandwidth extension is to minimize the high-band spectral distortion given the narrow-band spectral representative. The criterion is equal to minimize the mean square error of the LSF parameters in high-band.

$$min[d(k)] = min\{E[\|y(k) - \hat{y}(k)\|^2 \, |x(k)]\} \tag{79}$$

Where x and y are the LSF parameters in narrow-band and high-band [54, 55, **?**].

In conventional codebook mapping scheme, the minimization is implemented in two steps. First, the entries of the primary codebook C_x is defined by the vector quantization (VQ) of the speech feature $x(k)$. The quantization mapping Q is defined such as to minimize the criterion function $d(x(k), \hat{x}_i(k))$ between the input vector $x(k)$ and the entries $\hat{x}_i(k)$,

$$Q_s(x(k)) = \underset{\hat{x}_i(k) \in C_{x(k)}}{\arg\min} \; d(x(k), \hat{x}_i(k)) \tag{80}$$

where the distance measure $d(x(k), \hat{x}_i(k))$ is defined as Euclidian distance

$$d(x, \hat{x}_i) = \|x - \hat{x}_i\|^2 \tag{81}$$

Then, the corresponding entry in the shadow codebook is defined as:

$$\hat{y}_i = \underset{\hat{y}}{\arg\min} \, E\{d(y, \hat{y}_i) | x \in \gamma_i\} \tag{82}$$

where γ_i is the quantizer cell assigned to code vector $\hat{x}_i$. This equation is solved by the conditional expectation $\hat{y}_i = E\{y|x \in \gamma_i\}$. The expectation can be determined using a large number of pairs of training vectors $\{x(m), y(m)\}, m = 1...N_m$ by averaging the vectors y extracted from those signal frames for which $x(m) \in \gamma_i$.

It is well-known that each phoneme class has distinctive acoustic properties, thus distinctive LSF parameters in narrow-band and high-band. It means that better estimation of high-band LSF parameters can be achieved if the phoneme class of the signal in current frame is given. Therefore, equation 79 can be replaced by:

$$min[d(k)] = min\{E[\|y(k) - \hat{y}(k)\|^2 |x(k), x(k) \in \vartheta(k)]\} \tag{83}$$

Where $\vartheta(k)$ represents the phonetic label in the kth frame.

Generally, the codebook mapping and segmentation algorithm are computation complex. So, it is not practical to incorporate a segmentation algorithm in bandwidth extension. The other problems of using the segmentation are the variation of the split codebooks and large knowledge to be trained. The performance highly depends on the phonetic classification rate. Many artifacts will be introduced due to the misclassification. Therefore, in the proposed algorithm, we modify the conventional codebook mapping scheme towards increased phonetic classification. The entry in the shadow codebook alternatively defined as:

$$\hat{y}_i = \arg\min_{\hat{y}} E\{d(\hat{y}, \hat{y}_i)|x \in \vartheta_i\} \tag{84}$$

In this case, the training of the primary codebook is optimized to the following criterions:

$$Q(x) = \arg\min_{\hat{x}_i \in C_x} d(x, \hat{x}_i) \tag{85}$$

$$Q(x) = \arg\max_{\hat{x}_i \in C_x} P(\hat{x}_i \in \vartheta_i | x \in \vartheta_i, x \in \gamma_i) \tag{86}$$

There is no explicit method to optimize both criterions. In [92], the mutual information (/bit) between the short-time critical-band logarithm spectral energy and phonetic classification was found. The measure is for narrow-band signal, as shown in figure 31. Since a shift made on one LSF parameter is only related to local spectral distortion, thus change

the logarithm spectral energy in the corresponding critical-band. This information can be employed in the training of the primary codebook training, in order to increase the phonetic classification. Mutual information between data and phonetic labels is defined as:

$$I(x;L) = \int p(x,L) \log \frac{p(x,L)}{p(x)p(L)} dx dL \tag{87}$$

From the experimental results, we notice that the different frequency bands carry different information about phonetic classification. This distinction can be applied on vector quantization in the primary codebook. The alternative objective function is defined in the our system,

$$Q(x) = \arg\min_{\hat{x}_i \in C_x} \hat{d}(x, \hat{x}_i) \tag{88}$$

where the distance function is defined as:

$$\hat{d}(x, \hat{x}_i) = (x - \hat{x}_i)^T W (x - \hat{x}_i) \tag{89}$$

$$W = 2^{I(x \in \kappa, L)} \tag{90}$$

The parameter N is the dimension of the vector x, the weights w are illuminated in figure 31.

To evaluate the proposed distance measure, we define the hit rate of phonetic classification of a quantized cell γ_i as the maximal phonetic probability associated.

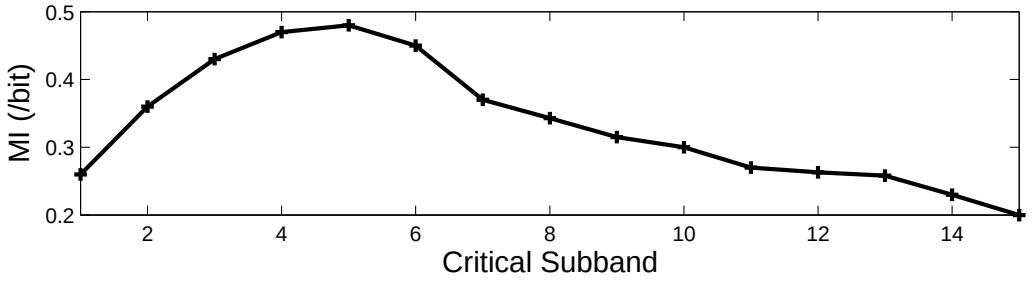
$$P(\gamma_i) = \arg\max_{\vartheta_i} P(\hat{x}_i \in \vartheta_i | x \in \vartheta_i, x \in \gamma_i) \tag{91}$$

The average hit rates in the overall primary codebook are measured and shown in table 14.

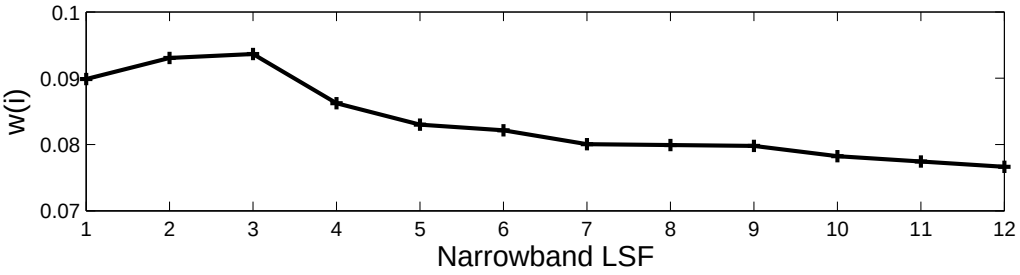
Table 14: The hit rate of phonetic classification based on codebook mapping

	Conventional	IPC
vowel	64.23%	**67.19**%
unvoiced fricative	54.47%	**61.01**%
nasal	62.73%	**70.44**%

The phonetic classification hit rates are increased by the codebook partition using the proposed distance function. The improved HB-LSD is also achieved, as indicated in table 15.



(a)



(b)

Figure 31: (a) The mutual information (/bit) between the short-time critical-band logarithm spectral energy and phonetic classification, (b) The proposed weighting for a distance measure toward increased phonetic classification.

Table 15: The performance of the improved codebook mapping towards increased phonetic classification (IPC)

	Conventional	IPC
HB-LSD (dB)	8.01	**7.12**

5.3.2 Marginal LSF interpolation

This algorithm makes the assumption that the frequency ranges $(0, \pi/2)$ and $(\pi/2, \pi)$ have the same number of LSF values in wideband speech. For example, when the LPC order is 24, there are always 12 LSFs in $(0, \pi/2)$ and 12 LSFs in $(\pi/2, \pi)$. This assumption is not true for all frames, and the actual distribution is (11,13) or (13,11) with a probability of around 50% in our training by TIMIT database. Particularly, the high-frequency consonants, which are of special interest to our problem, have mainly the distribution of (11,13). This reflects the fact that the speech energy is concentrated in the high-band. To compensate for this drawback, the proposed algorithm uses marginal LSF interpolation.

In the shadow codebook, the upper 13th-order LSF (12-24) is used instead of the upper 12th-order LSF (13-24). The mean of the lowest LSF in shadow codebook and the highest LSF in primary codebook will replace the highest LSF of actual narrowband speech in synthesis. The method is effective to lower the spectral distortion in the frequency region of (3KHz-5KHz).

Table 16: The performance of the marginal LSF interpolation

	Non-overlap codebook	Marg. LSF interp.
HB-LSD (dB)	7.12	**7.01**

5.3.3 Codebook Mapping With Memory

The conventional codebook mapping scheme is memoryless. The estimation of high-band LSF is based on the current time frame only. From the speech production model, there is correlation between each frame. Therefore, utilizing time history information can enable a more accurate codebook mapping.

$$min\{E[\|y(k) - \hat{y}(k)\|^2 \,|x(k), x(k-1), \hat{y}(k-1), x(k) \in \vartheta(k)]\} \tag{92}$$

We update the estimation of high-band LSF based on the interpolation of previous high-band LSF using the distance between their narrowband LSFs.

$$\hat{y}(k) = \beta \cdot \hat{y}(k) + (1-\beta) \cdot \hat{y}(k-1) \tag{93}$$

where, β is less than 1.

Table 17: The performance of codebook mapping with memory

	Memoryless	With memory
HB-LSD (dB)	7.01	**6.79**

5.3.4 Codebook Interpolation

It is shown, instead of a simple table lookup, the estimate $\hat{y}(k)$ is determined by a weighted sum of the most probable codebook entries.

$$\hat{y}(k) = \sum_{m=1}^{M} \alpha_m \hat{y}_m(k) \tag{94}$$

The individual weights α_m are inverse portional to the distance of the narrowband feature vector $x(k)$ to the respective mth closed primary codebook entry $\hat{x}_m(k)$.

$$\alpha_m = \frac{\frac{1}{\|x(k), x_m(k)\|^2}}{\sum_{m=1}^{M} \frac{1}{\|x(k), x_m(k)\|^2}} \tag{95}$$

Table 18: The performance of codebook mapping using interpolation

	NO interpolation	interpolated
HB-LSD (dB)	6.79	**6.63**

5.4 Performance Evaluation

Speech bandwidth extension is challenging in noisy environments, because a speech bandwidth extension system requires accurate estimation of narrowband spectral parameters. If the narrowband speech is corrupted by background noise, the estimation error in narrowband will introduce large error in the estimation in high-band. In this thesis, besides the performance evaluation in the Table. 19. in clean condition, we conducted experiments on noisy speech. The evaluation is important for predicting the performance in the real-world applications of a speech bandwidth extension system.

Table 19: The performance of the proposed bandwidth extension system (IPC) in clean conditions in term of log spectral distortion in high-frequency (4-8 KHz)

	Conventional	IPC	Marg. LSF Inter.	Memory Inter.	Multi Codewords
HB-LSD (dB)	8.01	7.12	7.01	6.79	6.63

5.4.1 Speech Spectrogram

Results of bandwidth extension of telephony speech are provided in Figure. 32, Figure. 33 and 34. Original speech with 8KHz bandwidth was downsampled to the bandwidth of 4KHz. Outputs were collected from the bandwidth extension mechanism.

The spectrograms indicate the outputs of the BWE reconstruct the missing high-frequency components.

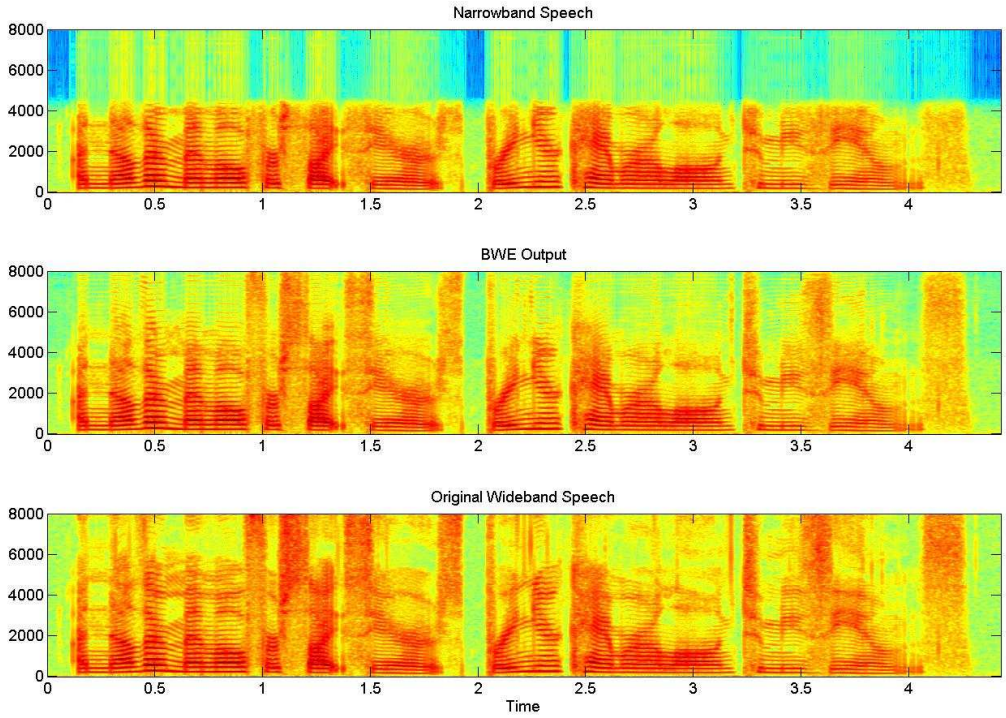


Figure 32: Speech spectrograms from bandwidth extension algorithm, (top) narrowband speech input with 4KHz bandwidth, (middle) output of bandwidth extension with 8KHz bandwidth, (bottom) original wideband speech

5.4.2 Objective Measurement

Our experiments were conducted in various additive noise conditions. The level of noise degradation is characterized by the narrowband LSD (NB-LSD). Two common noise types are used: room noise and car noise. The noisy signal is first processed by a noise suppressor. Then, the enhanced narrow speech is used for extending speech bandwidth. The overall LSD relative to clean speech is measured to indicate the performance improvement, as shown in

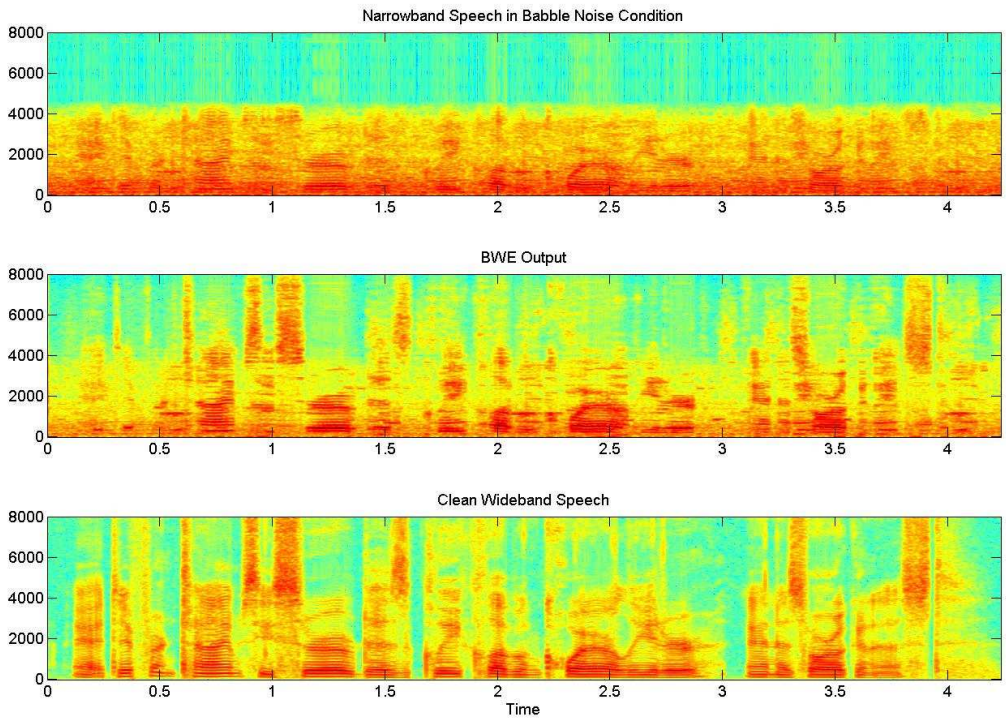


Figure 33: Speech spectrograms from bandwidth extension algorithm in babble noise condition, (top) noisy narrowband speech input with 4KHz bandwidth, (middle) output of bandwidth extension with 8KHz bandwidth, (bottom) clean wideband speech

Table 20. "CM" denotes the conventional codebook mapping method.

Table 20: The performance of the speech bandwidth extension in noisy conditions in terms of the narrowband LSD (NB-LSD) and Overall LSD relative to the original clean wideband speech

Noise Type	NB-LSD (dB)	Overall LSD (dB)		
		Noisy	CM	CM-IPC
Room	1.12	11.87	6.48	**5.43**
	2.06	12.35	7.07	**5.89**
	3.07	13.41	7.91	**6.72**
Car	1.07	11.23	6.28	**5.40**
	2.01	12.02	6.94	**5.87**
	3.11	13.35	7.87	**6.65**

5.4.3 Subjective Measurement

Mean Opinion Score (MOS) is probably the most widely used and simplest method to evaluate speech quality in general. It is also suitable for overall evaluation of synthetic

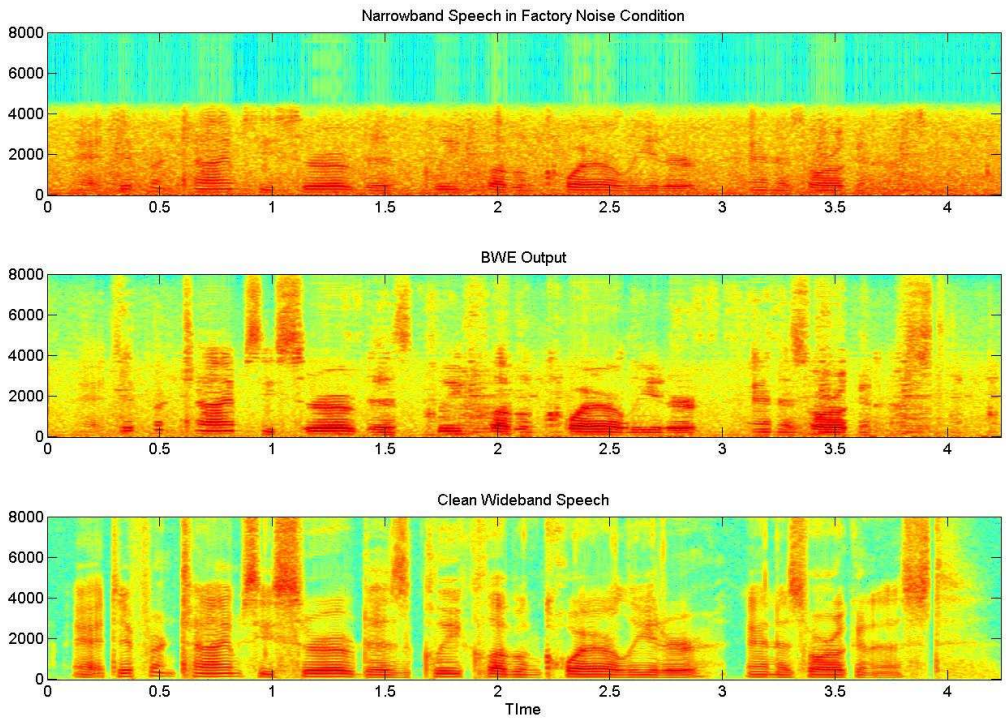


Figure 34: Speech spectrograms from bandwidth extension algorithm in factory noise condition, (top) noisy narrowband speech input with 4KHz bandwidth, (middle) output of bandwidth extension with 8KHz bandwidth, (bottom) clean wideband speech

speech. MOS is a five level scale from bad (1) to excellent (5) and it is also know as ACR (Absolute Category Rating). The listener's task is simply to evaluate the tested speech with scale described in Table 1.

In the test, the MOS test is run for the BWE outputs and narrowband speech. Table 21 lists the scores in clean and various noisy conditions. The speech data is selected from the TIMIT test sets "dr1" and "dr2". Noisy speech is acquired with additive noise. In all the data, 60% of sentences are female sentences, since the sounds from females are sharper and the high-frequency components are more important in perception.

Twenty listeners were involved in the test. Twelve of them are native English speakers. The baseline score "5" is given from the comparison between the narrowband and wideband speech, which referred to the enhanced outputs from the noise suppressor in noisy conditions. From the results in Table. 21 and Table. 22, the proposed bandwidth extension technique improves the speech quality. The improvement drops when noisy is added in, since

there are distortions in narrowband LSF estimation for this case. However, reconstructed high-frequency components boost the perceptual quality even introducing mis-estimation in certain frames. This is even more true for non-native speakers.

Table 21: MOS test for BWE outputs and narrowband speech on TIMIT sentences for native speakers

	Clean (BWE)	Car Noise (BWE+NS)	Babble Noise (BWE+NS)	Pink Noise (BWE+NS)	Factory Noise (BWE+NS)	White Noise (BWE+NS)
MOS	4.07	3.96	3.64	3.61	3.84	3.72

Table 22: MOS test for BWE outputs and narrowband speech on TIMIT sentences for non-native speakers

	Clean (BWE)	Car Noise (BWE+NS)	Babble Noise (BWE+NS)	Pink Noise (BWE+NS)	Factory Noise (BWE+NS)	White Noise (BWE+NS)
MOS	4.25	4.08	3.86	3.89	3.96	3.84

5.5 Summary

In this chapter, we proposed a speech bandwidth extension system by improved codebook mapping towards phonetic classification (CM-IPC). The model is simplified using a weighted distance function based on the mutual information between frequency bands and phonetic labeling. A set of techniques are used to improved the performance by marginal LSF interpolation, codebook mapping with memory, and codebook interpolation. The proposed system is effective in reducing the spectral distortion, thus helping to increase the objective quality of speech.

CHAPTER 6

CONCLUSION AND FUTURE WORKS

This thesis presents frameworks of speech enhancement in degraded environments of noise corruption or bandwidth limited transmission. The proposed algorithms include single-channel noise suppression using biologically inspired techniques, multi-sensor noise suppression for harsh environments using non-air conductive sensors, and speech bandwidth extension using codebook mapping toward increased phonetic classification.

The single-channel noise suppression algorithm conducts experimental trials for audio noise suppression modeling speech the perception mechanism of the human auditory system. Perceptually criterion and phoneme adaptive mechanism are imposed on the noise suppressor. The amount of suppression of this multiple–state algorithm is a non-linear function of the detected saliency and corresponding phoneme class. The algorithm effectively removes background noise. Although contained some residual noise, the enhanced outputs suffer less spectral distortion. Significant improvement on both speech quality and intelligibility were achieved. The potential of more refined phoneme adaptive strategies and robust functions to segmentation errors can be further worked on. And better results can certainly be obtained.

Classical speech enhancement algorithms utilize the inputs from acoustic sensors only. Although many optimization techniques have been developed, the enhanced outputs may still suffer from considerable speech distortion, especially in low SNR conditions. As a result, the speech intelligibility is degraded. In the proposed system, we exploit the state-of-the-arts non-acoustic sensors and introduce a hybrid speech enhancement system using perceptually inspired techniques. Both subjective and objective experiments were conducted in various noise types and levels. The proposed multi-sensor system is effective in suppressing background noise. Significant improvement of speech intelligibility for low-bit-rate speech coding in harsh environments is achieved.

The speech bandwidth extension system is based on the improved codebook mapping towards phonetic classification (CM-IPC). The model is simplified using a weighted distance function based on the mutual information between frequency bands and phonetic labeling. A set of techniques are used to improved the performance by marginal LSF interpolation, codebook mapping with memory, and codebook interpolation. The proposed system is effective in reducing the spectral distortion, thus help to increase the objective quality of speech.

REFERENCES

[1] Avendano, C., Hermansky, H., and Wan, E., "Beyond nyquist: towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proceedings of the EUROSPEECH*, vol. 1, (Madrid, Spain), pp. 165–168, Sept. 1995.

[2] Baken, R., "Electroglottography," in *J. Voice*, pp. 98–110, 1992.

[3] Bradley, J., "Speech intelligibility studies in classrooms," in *Journal of the Acoustical Society of America*, vol. 80, p. 846854, 1986.

[4] Brillinger, D. R., *Time Series: Date Analysis and Theory*. McGraw-Hill, New York, 1981.

[5] Burnett, G. C., Holzrichter, J. F., Gable, T. J., and Ng, L. C., "The use of glottal electromagnetic micropower sensors (GEMS) in determining a voiced excitation function." presented at the 138th Meeting of the Acoustical Society of America, Columbus, Ohio, Nov. 1999.

[6] Cappe, O., "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," in *IEEE Trans. Speech and Audio Proc.*, pp. 345 –349, apr 1994.

[7] Carl, H. and Heute, U., "Bandwidth enhancement of narrow-band speech signals," in *Proceedings of the EUSIPCO*, vol. 2, (Edinburgh, Scotland), pp. 1178–1181, Sept. 1994.

[8] Carlson, R., Granstrm, B., and Nord, L., "Evaluation and development of the kth text-to-speech system on the segmental level," in *Proceedings of ICASSP*, (Montreal, Canada), pp. 317–320, May 1990.

[9] Chen, G. and Parsa, V., "Hmm-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Proceedings of ICASSP*, (Montreal, Canada), pp. 709–712, May 2004.

[10] Cheng, Y., O'Shaughnessy, D., and Mermelstein, P., "Statistical recovery of wideband speech from narrowband speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 544–548, Oct. 1994.

[11] Chennoukh, S., Gerrits, A., Miet, G., and Sluijter, R., "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Proceedings of ICASSP*, pp. 665–668, May 2001.

[12] Cohen, I., "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 466–475, Sept. 2003.

[13] Cohen, I. and Berdugo, B., "Spectral enhancement by tracking speech presence probability in subbands," in *IEEE Workshop on Hands Free Speech Communication, HSC01*, (Kyoto, Japan), Apr. 2001.

[14] Cohen, I. and Berdugo, B., "Speech enhancement for nonstationary noise environments," *Signal Processing*, vol. 81, p. 24032418, Nov. 2001.

[15] Cover, T. M. and Thomas, J. A., *Elements of Information Theory*. Wiley, New York.

[16] Darbellay, G. and Vajda, I., "Estimation of the information by an adaptive partition of the observation space," *IEEE Transaction on Information Theory*, vol. 45, pp. 1315–1321, May 1999.

[17] Doblinger, G., "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *4th Europe Conference on Speech, Communication, and Technology, EUROSPEECH95*, (Madrid, Spain), p. 15131516, Sept. 1995.

[18] Drucker, H., "Speech processing in a high ambient noise environment," *Audio and Electroacoustics, IEEE Transactions on*, vol. 16, pp. 165–168, June 1968.

[19] Enborn, N. and Kleijn, W. B., "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Proceedings of the IEEE Workshop on Speech Coding*, (Porvoo, Finland), pp. 171–173, Sept. 1999.

[20] Ephraim, Y., "Statistical model based speech enhancement systems," *Proceeding IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[21] Ephraim, Y. and Malah, D., "Speech enhancement using optimal non-linear spectral amplitude estimation," in *Proceedings of ICASSP*, (Boston, USA), pp. 1118–1123, 1983.

[22] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transaction on Acoustic, Speech, Signal Processing*, vol. ASSP-32, p. 11091121, Dec. 1984.

[23] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Transaction on Acoustic, Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[24] Ephraim, Y. and Trees, H. L. V., "signal subspace approach for speech enhancement," *IEEE Transaction on Acoustic, Speech, Signal Processing*, vol. 3, pp. 251–266, July 1995.

[25] Ephraim, Y. and Trees, H. L. V., "signal subspace approach for speech enhancement," *IEEE Transaction of Information Theory*, vol. 48, pp. 1518–1569, June 2002.

[26] Epps, J. and Holmes, W., "A new technique for wideband enhancement of coded narrowband speech," in *Proceedings of the IEEE Workshop on Speech Coding*, (Porvoo, Finland), pp. 174–176, Sept. 1999.

[27] Ertan, A. and III, T. B., "Circular linear predictive modeling for speech coding application," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov. 2003.

[28] Fuemmeler, J. A., Hardie, R. C., and Gardner, W. R., "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 2001 n.4, pp. 266–274, Dec. 2001.

[29] G.712, I.-T. R., "Transmission performance characteristics of pulse code modulation channels," *ITU-T Rec.G712, International Telecommunications Union*, 1986.

[30] G726, I.-T. R., "40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (adpcm) 1951," *ITU-T Rec.G726 Document Number E, International Telecommunications Union*, 1991.

[31] G729, I.-T. R., "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (cs-acelp)," *ITU-T Rec.G729 Document Number E 10204, International Telecommunications Union*, 1996.

[32] Ganapsthiraju, A., Hammake, J. E., and Picone, J., "Signal modeling techniques in speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 81, pp. 1215–1247, Sept. 1993.

[33] Gazor, S. and Zhang, W., "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 498–505, Sept. 2003.

[34] Goldstein, M., "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Communication*, vol. 16, pp. 225–244, 1995.

[35] Gustafsson, H., Claesson, I., and Lindgren, U., "Speech bandwidth extension," in *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, pp. 809–812, Aug. 2001.

[36] Hansen, J. and Clements, M., "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, pp. 795–805, 1991.

[37] Harma, A., "Temporal masking effects: Single incidents." FAMbac Technical Report, 1999.

[38] Heide, D. and Kang, G., "Speech enhancement for bandlimited speech," in *Proceedings of ICASSP*, vol. 1, (Seattle, USA), pp. 393–396, May 1998.

[39] Hirsch, H. G. and Ehrlicher, C., "Noise estimation techniques for robust speech recognition," in *Proceedings of ICASSP*, (Detroit, USA), p. 153156, May 1995.

[40] Hu, R. and Anderson, D. V., "Audio noise suppression based on neuromorphic saliency and phoneme adaptive filtering," in *IEEE DSP Workshop*, (Taos, NM), Aug. 2004.

[41] Hu, R. and Anderson, D. V., "Improved perceptually inspired speech enhancement using an auditory model," in *Proceedings of the Asilomar Conference on Circuits, Systems, and Computers*, Nov. 2004.

[42] Hu, R. and Anderson, D. V., "Single acoustic channel speech enhancement based on glottal correlation using non-acoustic sensors," in *International Conference on Spoken Language Processing*, (Jeju, Korea), Oct. 2004.

[43] Hu, R. and B.Raj, "A robust voice activity detector using an acoustic doppler radar," in *2005 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU2005)*, (Cancun, Mexico), Dec. 2005.

[44] Hu, R., Kamath, S., and Anderson, D. V., "Speech enhancement using non-acoustic sensors," in *Europe Conference on Speech Communication Technology*, (Lisbon, Portugal), Sept. 2005.

[45] Hu, R., Krishnan, V., and Anderson, D. V., "Speech bandwidth extension by codebook mapping towards increased phonetic classification," in *Europe Conference on Speech Communication Technology*, (Lisbon, Portugal), Sept. 2005.

[46] Jax, P., *Enhancement of Bandlimited Speech Signals: Algorithms ans Theretical Bounds*. Ph.D Thesis, RWTH Aachen, 2002.

[47] Jax, P. and Vary, P., "Wideband extension of telephone speech using a hidden markov model," in *Proceedings of the IEEE Workshop on Speech Coding*, (Delavan, USA), pp. 133–135, Sept. 2000.

[48] Jax, P. and Vary, P., "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov mode," in *Proceedings of ICASSP*, vol. 1, pp. 680–683, 2003.

[49] Jax, P. and Vary, P., "On artificial bandwidth extension of telephone speech," *IEEE Transactions on Signal Processing*, vol. 83, pp. 1707–1719, Aug. 2003.

[50] Jax, P. and Vary, P., "Feature selection for improved bandwidth extension of speech signals," in *Proceedings of ICASSP*, (Montreal, Canada), pp. 697–700, May 2004.

[51] Jekosch, U., "Speech quality assessment and evaluation," in *Europe Conference on Speech Communication Technology*, pp. 1387–1394, 1993.

[52] Kim, W., Kang, S., and Ko, H., "Spectral substraction based on phonetic dependency and masking effects," in *IEE Proceeding on Visual and Image Process*, Oct. 2000.

[53] Kraft, V. and Portele, T., "Quality evaluation of five german speech synthesis systems," in *Acta Acustica*, 1995.

[54] Krishnan, V. and Anderson, D. V., "Robust jointly optimized multistage vector quantization for speech coding," in *eurospeech*, 2003.

[55] Krishnan, V. and Anderson, D. V., "Joint design of channel-optimized multistage vector quantizers," in *IEEE Signal Processing Letters*, 2004.

[56] Larsen, E. and Aarts, R. M., *Audio Bandwidth Extension: Application of Pstchoacoustic, Signal Processing and Loudspeaker Design*. Whitley, 2004.

[57] Lim, J. and Oppenheim, A., "All pole modelling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, p. 197210, 1978.

[58] Lochner, J. and Burger, J., "The influence of reflections inauditorium acoustics," *Journal of Sound and Vibration*, vol. 4, p. 426454, 1964.

[59] Logan, J., Greene, B., and Pisoni, D., "Segmental intelligibility of synthetic speech produced by rule," *Journal of the Acoustical Society of America*, vol. 86, pp. 566–581, 1986.

[60] Makhoul, J. and Berouti, M., "High-frequency regeneration in speech coding systems," in *Proceedings of ICASSP*, vol. 4, (Washington, USA), pp. 428–431, Apr. 1979.

[61] Mariniak, A., "A global framework for the assessment of synthetic speech without subjects," in *Europe Conference on Speech Communication Technology*, pp. 1683–1686, 1993.

[62] Martin, R., "Spectral subtraction based on minimum statistics," *7th Europe Signal Processing Conference (EUSIPCO94)*, p. 11821185, Sept. 1994.

[63] Martin, R., "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 504–512, 2001.

[64] McAulay, R. J. and Malpass, M. L., "Speech enhancement using a softdecision noise suppression filter," in *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, p. 137145, Apr. 1980.

[65] McKinley, B. L. and Whipple, G. H., "Model based speech pause detection," in *Proceedings of ICASSP*, (Munich, Germany), p. 11791182, Apr. 1997.

[66] Messing, D., *Noise Suppression with Non-Air-Acoustic Sensors*. Masters Thesis, MIT, Sept. 2003.

[67] Meyer, J., Simmer, K. U., and Kammeyer, K. D., "Comparison of one- and two-channel noise-estimation techniques," in *5th International Workshop on Acoustic Echo and Noise Control (IWAENC97)*, (London, U.K.), p. 137145, Sept. 1997.

[68] Miet, G., Gerrits, A., and Valiere, J. C., "Low-band extension of telephone-band speech," in *Proceedings of ICASSP*, (Istanbul, Turkey), pp. 1851–1854, June 2000.

[69] Musicus, B. R., "An iterative technique for maximum likelihood parameter estimation on noisy data," in *S.M. Thesis, M.I.T.*, (Cambridge, Massachusetts), 1979.

[70] Musicus, B. R., "Maximum likelihood parameter estimation of noisy data," in *Proceedings of ICASSP*, 1979.

[71] Ng, L. C., Burnett, G. C., Holzrichter, J., and Gable, T. J., "Denoising of human speech using combined acoustic and em sensor signal processing," in *Proceedings of ICASSP*, (Istanbul, Turkey), June 2000.

[72] Ng, L., Burnett, G., Holzrichter, J., and Gable, T., "Background speaker noise removal using combined EM sensor/acoustic signal signals." presented at the 138th Meeting of the Acoustical Society of America, Columbus, Ohio, Nov. 1999.

[73] Nilsson, M., Andersen, S., and Kleijn, W., "On the mutual information between frequency bands in speech," in *Proceedings of ICASSP*, June 2000.

[74] Ochsman, R. and Chapanis, A., "The effects of 10 communication modes on the behaviour of teams during co-operative problem-solving," in *International Journal of Man-Machine Studies*, 1974.

[75] Park, K. and Kim, H., "Narrowband to wideband conversion of speech using gmm-based transformation," in *Proceedings of ICASSP*, vol. 3, (Istanbul, Turkey), pp. 1847–1850, June 2000.

[76] Picone, J., "Signal modeling techniques in speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 81, pp. 1215–1247, Sept. 1993.

[77] Priestley, M. B., *Spectral Analysis and Time Series*. Academic Press, 1989.

[78] Qian, Y. and Kabal, P., "Dual-mode wideband speech recovery from narrowband speech," in *Proceedings of EUROSPEECH*, pp. 1433–1437, Sept. 2003.

[79] Qian, Y. and Kabal, P., "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proceedings of ICASSP*, (Montreal, Canada), pp. 713–716, May 2004.

[80] Quatieri, T., *Discrete-Time Speech Signal Processing: Principles And Practice*. Prentice Hall PTR, 2001.

[81] Ris, C. and Dupont, S., "Assessing local noise level estimation methods: Application to noise robust asr," *Speech Communication*, vol. 34, p. 141158, Apr. 2001.

[82] Scanlon, M., "Acoustic sensor for health status monitoring," in *Proceeding of IRIS Acoustic and Seismic Sensing*, vol. II, pp. 205–222, 1998.

[83] Sinha, D. and Tewfik, A. H., "Low bit rate transparent audio compression using adapted wavelets," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3463–3479, Dec. 1993.

[84] Sohn, J., Kim, N. S., and Sung, W., "A statistical model-based voice activity detector," in *IEEE Signal Processing Letter*, vol. 6, pp. 1–3, May 1999.

[85] Sohn, J. and Sung, W., "A voice activity detector employing soft decision based noise spectrum adaption," in *Proceedings of ICASSP*, vol. 1, pp. 365–368, May 1998.

[86] Stahl, V., Fischer, A., and Bippus, R., "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proceedings of ICASSP*, vol. 3, pp. 1875–1878, June 2000.

[87] Tanyer, S. G. and Ozer, H., "Voice activity detection in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 478–482, July 2000.

[88] Unno, T. and McCree, A., "A robust narrowband to wideband extesion system featuring enhanced codebook mapping," in *Proceedings of ICASSP*, (Philadelphis, USA), pp. 805–808, Mar. 2005.

[89] Valin, J. M. and Lefebvre, R., "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *Proceedings of IEEE Workshop on Speech Coding*, (Delavan, USA), pp. 130–132, Sept. 2000.

[90] Vapnik, V., *Statistical Learning Theory*. New York: Wiley, 1998.

[91] Virag, N., "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.

[92] Yang, H. H., Sharma, S., van Vuuren, S., and Hermansky, H., "Relevance of timefrequency features for phonetic and speakerchannel classification," *Speech Communication*, vol. 31, pp. 35–50, Aug. 2000.

[93] Yang, W., *Enhanced Modified Bark Spectral Distortion (EMBSD): An ObjectiveE Speech Quality Measure Based On Audible Distortion and Cognition Model*. Ph.D Thesis, Temple University, May 1999.

VITA

Rongqiang Hu was born in 1979 at Beihai city, Guangxi province, China. At age of 18, he attended Beihang University, Beijing, China, as an undergraduate student, and finished a B.S. degree in July, 2001. After that time, he joined Georgia Institute of Technology for the graduate program in electrical and computer engineering, where he received a M.S. degree in August, 2002. Then, he continued the Ph.D program advised by Prof. David Anderson in cooperative analog and digital signal processing (CADSP) lab, a subgroup of center for signal and image processing (CSIP). During the summer of 2005, he had worked with Dr. Bhiksha Raj as an intern in MITSUBISHI Electric Research Laboratories (MERL), where he participated in the project on denoising speech using secondary sensors.

Rongqiang's research interests include general topics in digital signal processing, especially a variety of topics in speech and audio processing: noise suppression, echo cancelation, audio bandwidth extension, psychoacoustic modeling, speech coding, speech recognition, and audio-visual fusion.